Universidad Politécnica de Madrid

Facultad de Informática

UNIVERSIDAD POLITÉCNICA DE MADRID

FACULTAD DE INFORMÁTICA

TRABAJO FIN DE GRADO

Análisis de datos sobre virus Informáticos

Autor: Zhaneta Dinkova
Tutor: Aurora Pérez

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

I. AGRADECIMIENTOS

En primer lugar me gustaría dar las gracias a mi madre, a mi padre, a mi hermana y a toda mi familia por haberme apoyado y creído en mí a lo largo de estos seis años.

También quiero destacar a mi tutora, Aurora Pérez, por haberme dado esta oportunidad y por haberme ayudado y guiado en todo momento.

Me gustaría mencionar de forma muy especial a todos mis compañeros y amigos de la facultad. Muchos de ellos me han apoyado en momentos difíciles y me han dado fuerzas cuando más lo necesitada, y se lo agradezco de verdad.

Tampoco quiero dejar fuera a mis inseparables amigas, que como yo, saben lo que supone llegar hasta aquí.

Por último, quiero señalar a todos los profesores que he tenido durante todo este tiempo, a su labor y dedicación a nosotros.

II. RESUMEN

El trabajo fin de grado que se presenta en este documento trata de "*Aplicar técnicas de Data Mining a un conjunto de datos procedentes de ataques de virus informáticos interceptados en servidores de Internet*".

La propuesta de este trabajo surgió de una Institución con el fin de extraer información de un conjunto de datos proveniente de ejecuciones de virus informáticos. Lamentablemente, debido a fuertes restricciones de privacidad por parte de esta Institución y así como al relevo de la persona responsable de éste área en dicha Institución, el Proyecto finalmente se canceló.

Como consecuencia, y teniendo en cuenta el carácter didáctico de este trabajo fin de grado, el proyecto KDD (Knowledge Discovery in Databases) en sí y sus objetivos de negocio y objetivos de data mining, se han establecido conforme con la misma temática de predicción de ataques de virus que había planteado la Institución en el pasado, contando con una base de datos que ha sido recopilada de diferentes empresas anónimas.

Para llevar un desarrollo estructurado de todas las fases del proceso KDD, se ha trabajado siguiendo como referencia una metodología para proyectos de Data Mining, "*CRISP-DM*", cuyo estándar incluye un modelo y una guía, estructurados en seis fases.

Como herramienta de Data Mining a utilizar, se ha elegido el software de libre distribución "*WEKA*".

Por último, cabe destacar que el proyecto ha concluido satisfactoriamente, lográndose cada una de las metas establecidas como proyecto de minería de datos.

III. INTRODUCCIÓN y OBJETIVOS

Este trabajo consiste en aplicar técnicas de Data Mining a un conjunto de datos procedentes de ataques de virus informáticos interceptados en servidores de Internet.

Estaba previsto que este trabajo se enmarcara dentro de un Proyecto global de KDD con la Institución propietaria de los datos, con la que se habían iniciado conversaciones tiempo atrás, pero debido a fuertes restricciones de privacidad por parte del cliente, así como al relevo de la persona responsable del área en dicha Institución, el Proyecto se canceló. Por este motivo, el presente trabajo se ha realizado sin contar con el soporte del cliente para poder entender tanto los datos como el negocio. Pese a ello, se ha realizado el trabajo enfocando los objetivos de negocio y los de data mining en la misma temática de predicción de ataques de virus que había planteado el cliente en el pasado. Por otro lado, con la realización de este trabajo de minería de datos, se pretende profundizar y adquirir nuevos conocimientos en todo el proceso de Descubrimiento de Conocimiento en Bases de Datos, objetivo que ha sido plenamente logrado. Para seguir un desarrollo estructurado de Data Mining, se ha tomado como referencia la metodología "*CRISP-DM 1.0*" , un modelo jerárquico de libre distribución que proporciona una clara descripción del ciclo de vida de un proyecto de Data Mining.

En la siguiente *tabla* se pueden observar los objetivos de este trabajo. Estos objetivos son las metas que se pretenden conseguir con la realización de este Trabajo Fin de Grado para descubrir conocimiento nuevo en grandes volúmenes de conjuntos de datos.

Nº obj.	**Descripción del objetivo**
1	Definir el tipo de conocimiento que se espera encontrar
2	Definir el algoritmo de data mining a utilizar
3	Aprender a utilizar y conocer el proceso software "*CRISP-DM 1.0*"
4	Aprender a utilizar y conocer la herramienta "*WEKA*"
5	Realizar la selección de los datos
6	Realizar la preparación y limpieza de los datos
7	Realizar la transformación de los datos
8	Definir (o establecer) los parámetros con que se ejecutará el algoritmo de data mining a aplicar
9	Modelar los datos
10	Evaluar los resultados obtenidos

Tabla 1: Objetivos del trabajo

El orden de los objetivos 5 – 10 se rige por la evolución y fases que sigue el *proceso KDD* (Knowledge Discovery in Databases) propuesto por Osama Fayyad en 1996 tal como se puede observar en la *Figura **1***.

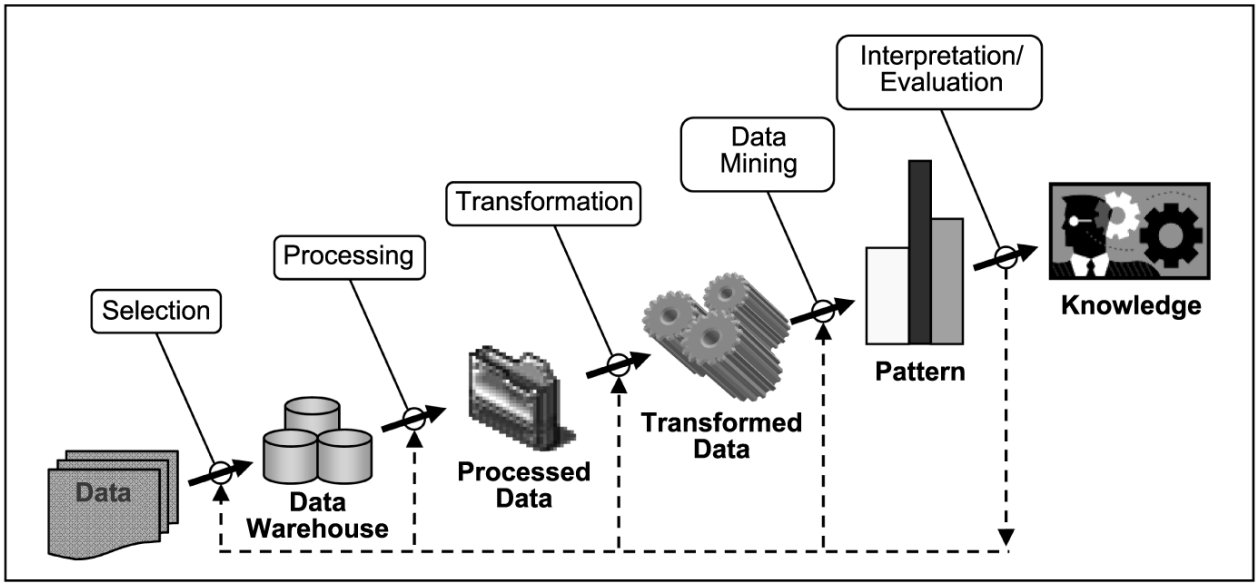

Figura 1:Fases de proceso KDD

Fases del proceso KDD

1.Selección de datos. En esta etapa se determinan las fuentes de datos a utilizar y el tipo de información a analizar.

2.Pre-procesamiento. Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de datos para ser manejados en las fases posteriores.

3.Transformación. Esta etapa consiste en el tratamiento preliminar de los datos, transformación y generación de nuevas variables a partir de las ya existentes con una estructura de datos apropiada.

4.Data Mining. Es la fase de modelado de los datos, donde se aplican diferentes técnicas con el objetivo de extraer patrones nuevos y potencialmente útiles que están contenidos u "ocultos" en los datos.

5.Interpretación y Evaluación. Se identifican los patrones obtenidos y se realiza una evaluación de los resultados obtenidos. Si los resultados no son buenos, se volverá a alguna de las fases anteriores.

IV. TRABAJOS PREVIOS

Son numerosas las aplicaciones de análisis de datos referentes a la seguridad informática: detección de intrusos, comportamientos anómalos, optimización de recursos de redes y servidores, detección de fraudes, etc.. , enfocándose en la *prevención* y *protección* de la infraestructura computacional y en la privacidad de los datos.

Trabajos como *Virus detection using data mining techniques,* de Jau-Hwang Wang et al**.** tratan de buscar, a través de técnicas de minería de datos, patrones que se encuentran en una colección de códigos de virus para posteriormente utilizarlos en la detección de nuevos ejecutables maliciosos. En concreto, este trabajo utiliza las técnicas de clasificación de data mining y el modelo probabilístico de redes bayesianas para extraer estos nuevos patrones.

Este es sólo un ejemplo de los diversos trabajos que existen en el campo de la seguridad informática que tratan de analizar diferentes bases de datos de virus aplicando distintas técnicas de data mining con el fin de encontrar información valiosa que pueda emplearse para proteger y prevenir.

La evaluación de los resultados obtenidos en la última fase del proceso KDD no siempre muestra un resultado bueno. En este caso, siendo KDD un proceso iterativo, se podrá reiterar en los pasos anteriores con el fin resolver conflictos y errores encontrados.

V. DESARROLLO

1. COMPRENSIÓN DEL PROBLEMA

1.1 Determinar los Objetivos del Negocio

Como se ha mencionado anteriormente, el presente trabajo se ha realizado sin contar con el soporte del cliente para poder comprender tanto los datos como el negocio. A pesar de ello, el trabajo se ha efectuado enfocando los objetivos de negocio y los de data mining en la misma temática de predicción de ataques de virus que había planteado el cliente en el pasado, y más concretamente, para el desarrollo de un programa software capaz de predecir los ataques que puedan sufrir las diferentes instituciones, permitiendo así tomar las medidas de prevención adecuadas contra estos virus.

Para lograr el objetivo de prevenir los ataques en la red, se han fijado los siguientes objetivos de negocio:

N° obj.	Objetivo
1	Conocer características de todos los virus interceptados con el fin de predecirlos.
2	Conocer información relevante sobre los virus que atacan a cada sector de institución para poder tomar medidas de prevención frente a ellos.

Tabla 2: Objetivos del negocio

1.1.1 Criterios de Éxito del Negocio

En relación a cada objetivo de negocio se establece un criterio de éxito que lo satisfaga.

N°	Medida de éxito
1	Obtener información útil para poder prevenir los ataques.
2	Obtener información relevante de los virus que atacan a cada ámbito de institución.

Tabla 3:Medidas de éxito

1.2 Valoración de la situación

1.2.1 Inventario de recursos

Roles

	Rol	Funciones
1	Jefe de proyecto	Responsable de la planificación del proyecto. Dirección y coordinación de los recursos empleados. Propone, en su caso, modificaciones en la ejecución del proyecto.
2	Analista de minería de datos	Responsable de transformar los objetivos del negocio en objetivos de Data Mining (DM). Realiza todas las fases del proceso DM: ▪ Define el(los) algoritmo(s) DM a utilizar ▪ Realiza la selección de los datos ▪ Realiza la preparación y limpieza de los datos ▪ Realiza la transformación de los datos ▪ define (o establece) los parámetros del algoritmo DM que se va a aplicar ▪ Ejecuta el(los) algoritmo(s) DM ▪ Evalúa los resultados obtenidos ▪ Transforma los resultados en información interpretable por el usuario final

Tabla 4: Roles

Persona asociado a cada rol

	Rol	Nombre	Disponibilidad
1	Jefe de proyecto	Aurora Pérez	Consultar por email aurora@fi.upm.es
2	Analista de minería de datos	Zhaneta Dinkova	Consultar por email z.dinkova@alumnos.upm.es

Tabla 5: Persona asociado a cada rol

Recursos Hardware

	Hardware	Descripción
1	MacBook Pro 5	Intel Core 2 Duo2,53 GHz 4GB

Tabla 6: Hardware

Recurso Software

	Herramienta Data Mining	Versión
1	WEKA[1]	3.6.9

Tabla 7: Software

Fuentes de datos

id	Nombre	Registros	Tamaño
1	T_ALIAS	14.335	674.5 Kb
2	T_VIRUS	6.347	27.0 MB
3	inci_virus	1.801.886	152.9 MB
4	inci_informes	68.371	12.3 MB
5	inci_sensores	82	24.2 MB
6	inci_ambitos	7	7 MB

Tabla 8: Fuentes de datos

[1]**Weka** (Waikato Environment for KnowledgeAnalysis - *Entorno para Análisis del Conocimiento de la Universidad de Waikato) es una plataforma de software para aprendizaje automático escrito en Java y desarrollado en la Universidad de Waikato.*

1.2.2 Requerimientos y restricciones

1.2.2.1 Requisitos del proyecto

Id	Descripción
1	Presentar un documento final en el que se explica todo el proceso y resultados obtenidos.
2	Fecha de entrega: 9 de junio a las 23:55 horas.
3	324 horas de trabajo para la realización del proyecto.

Tabla 9: Requisitos

1.2.2.3 Restricciones

Id	Descripción
1	La colección de datos es facilitada en formato CSV, delimitado por comas.
2	No se dispone de persona de contacto para posibles consultas sobre la colección de datos.

Tabla 10: Restricciones

1.2.3 Riesgos y Contingencias

1.2.3.1 Plan de riesgos y contingencia

El presente plan identifica posibles accidentes (riesgos) que pueden ocurrir durante la realización del trabajo y responde a los mismos con una medida de contingencia, garantizando así la continuidad del negocio.

Id	Riesgo	Probabilidad	Impacto	Mitigación del riesgo
1	No identificar alguno de los objetivos del negocio o de Data Mining.	Baja	Alto	Incluir en el proyecto toda la información relativa al nuevo objetivo.
2	Identificar algún objetivo de negocio o de Data Mining que no es realmente un objetivo.	Baja	Alto	Empezar el proyecto de nuevo. Dedicar más tiempo a esta tarea para evitar el mismo error.
3	Obtener resultados incorrectos.	Media	Alto	Repetir algunos pasos para obtener resultados correctos. Poner hincapié en los pasos y en la evaluación de resultados para evitar el mismo error.
4	No detectar resultados incorrectos[2]	Baja	Alto	Empezar el proyecto de nuevo. Contratar nuevos especialistas que desempeñen el proyecto.
5	No cumplir con el plazo de entrega.	Baja	Alto	Trabajar más horas de las planificadas, incluyendo horario fuera de lo establecido.
6	Otros trabajos influyen en la planificación realizada.	Media	Medio	Reestructurar el horario y planificarlo en función a la disponibilidad del miembro.
7	Baja de algún miembro del proyecto	Alta	Alto	Reestructurar el horario en función de la disponibilidad de los miembros del proyecto.
8	Trabajar más horas de lo planificado.	Baja	Bajo	Mayor coste del proyecto. Reestructurar las tareas de cada miembro.

[2] Se entiende que este riesgo es improbable que ocurra puesto que el proyecto se realiza por analistas expertos en el tema.

Id	Riesgo	Probabilidad	Impacto	Mitigación del riesgo
9	Trabajar menos horas de lo planificado	Baja	Medio	Reestructurar el reparto de tareas. Asignar más tareas a un miembro o aumentar sus horas de trabajo.
10	Perder el trabajo hecho hasta ahora.	Media	Alto	Empezar el trabajo desde la última versión guardada en el servidor*backup*.
11	No poder recuperar algún fichero *backup*	Baja	Alto	Empezar el trabajo desde la última versión guardada en el servidor*backup*.Si no se dispone de ninguna versión, empezar el proyecto de cero.

Tabla 11: Plan de riesgos y contingencia

1.2.4 Terminología

1.2.4.1 Terminología específica del negocio

Nombre	**Descripción**
CRISP_DM	*CRISP–DM es una metodología para el desarrollo de proyectos de Data Mining, estructura en seis diferentes fases.*
Backup	*Es una copia de los datos originales que se realiza con el fin de disponer de un medio de recuperarlos en caso de su pérdida*
Ámbito	*Es el sector en que se agrupan las instituciones*
Sensor	*Es la institución, empresa.*

Tabla 12: Terminología de negocio

1.2.4.2 Terminología especifica de Data Mining

Nombre	***Descripción***
Algoritmo	*Conjunto finito de instrucciones o pasos que sirven para ejecutar una tarea o resolver un problema*
Centroide	*Utilizado en el proceso de agrupamiento k-medias, es el centro de los grupos de agrupamiento.*
Clasificación	*Es la acción o el efecto de ordenar o disponer por clases*
Clasificación	*Es la acción o el efecto de ordenar o disponer por clases.*
Cobertura	*La cobertura mide la proporción de términos correctamente reconocidos respecto al total de términos reales.*
Conjunto de entrenamiento	*Conjunto de datos que se utiliza para construir el modelo*
Conjunto de testeo o de prueba	*Conjunto de datos de prueba que se utilizan para probar el modelo obtenido con el conjunto de datos de entrenamiento.*
Data Mining	*Proceso no trivial de extracción de conocimiento útil y previamente desconocido a partir de un conjunto de datos*
DM	*Siglas de Data Mining.*
Log likelihood	*Logaritmo de la verosimilitud. En estadística, la estimación por máxima verosimilitud es un método habitual para ajustar un modelo y encontrar sus parámetros. Su valor es siempre menor que 1 y por lo tanto su logaritmo será negativo.*
Matriz de confusión	*Es una herramienta de visualización que se emplea en aprendizaje superficial. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real.*
Missing	*Valor en blanco, indefinido en un campo de un registro.*
Moda	*En estadística, la moda es el valor con una mayor frecuencia en una distribución de datos.*
Null	*Ausencia de valor en un campo de un registro.*
Precisión	*Mide el número de términos correctamente reconocidos respecto al total de términos predichos.*
Ratio de ganancia	*El ratio de ganancia de un atributo es la relación entre la ganancia de información y su valor intrínseco. Se utiliza para evitar considerar atributos con un gran número de valores distintos.*

Nombre	***Descripción***
Variable cualitativa Nominal	*Presenta modalidades no numéricas que no admiten un criterio de orden*
Variable cualitativa Ordinal	*Presenta modalidades no numéricas en las que existe un orden*
Variable cuantitativa Continua	*Puede adquirir cualquier valor dentro de un intervalo especificado de valores.*
Variable cuantitativa Discreta	*No admite valores intermedios entre dos valores específicos*
Variable predictora	*Es un atributo que se utiliza en la construcción del modelo clasificador y se selecciona cuando mejor separa los ejemplos de acuerdo con la clase seleccionada.*

Tabla 13: Terminología de Data Mining

1.2.5 Costes y Beneficios

No se especificarán los costes ni los beneficios del proyecto, ya que realmente no se dispone de un cliente final para el cual realizar estas estimaciones.

1.3 Objetivos Data Mining

Nº	Objetivo DM	Descripción	Objetivo de Negocio
1	Obtener una clasificación para la clase "tipo" de virus	Obtener un modelo de clasificación utilizando como variable a predecir la variable "tipo" de la tabla T_VIRUS.	nº1 Conocer características de todos los virus interceptados con el fin de predecirlos.
2	Obtener una clasificación para una clase "cluster" previamente generada	Obtener un modelo de clasificación utilizando como variable a predecir la clase "cluster" que se obtendrá previamente aplicando una técnica de *clustering*. Se trabajará con los datos de la tabla T_VIRUS.	
3	Obtener una clasificación para la clase "id_ámbito"	Obtener un modelo de clasificación utilizando como variable a predecir la variable "id_ambito", que se corresponde con el sector al que pertenece la institución. Dicha variable "id_ambito" se obtendrá combinando información de las tablas T_VIRUS, inci_virus e inci_sensores.	nº2 Conocer información relevante sobre los virus que atacan a cada ámbito (sector en que se agrupan las instituciones), para así poder tomar medidas de prevención frente a ellos.

Tabla 14: Objetivos de Data Mining

La técnica y pasos que se han utilizado para obtener la clases "cluster" a predecir en el objetivos nº 2 se explicará en el correspondiente apartado de *Modelado*, y para la clase id_ambito, necesaria para el objetivo 3 de data mining, en el apartado 3.4 Integración de los datos. La variable *cluster* se obtendrá tras realizar *clustering* sobre tres de los atributos de la tabla T_VIRUS y la variable id_ambito, se obtendrá combinando varias tablas (T_VIRUS, inci_virus e inci_sensores).

Se ha intentado que los objetivos de Data Mining sean los más apropiados para cada uno de los objetivos de negocio establecidos. Sin embargo, tal como se ha explicado anteriormente, estos objetivos son ficticios ya que, lamentablemente, las altas restricciones por parte del cliente impidieron que se firmara el contrato para la realización del proyecto.

Para el objetivo de negocio nº1 "*Conocer características de todos los virus interceptados con el fin de predecirlos*" se han planificado dos objetivos de Data Mining:

i. Objetivo de Data Mining 1
Para alcanzar este objetivo se construirá un modelo de clasificación según la variable "tipo", es decir, se realizará una clasificación por tipo de virus, consiguiendo descubrir una serie de reglas que se cumplen para cada tipo de virus.

ii. Objetivo de Data Mining 2
Para este segundo objetivo, primero se utilizará la técnica de agrupamiento para establecer el número de clusters en que se dividen todas las instancias, agrupando en cada uno de ellos las instancias que más se asemejan entre sí. De esta manera, la clasificación posterior se realizará sobre el conjunto de clusters obtenido, y mostrará las características que tienen los virus que pertenecen al mismo cluster.

Para el segundo objetivo de negocio "*Conocer información relevante sobre los virus que atacan a cada ámbito (sector en que se agrupan las instituciones), para así poder tomar medidas de prevención frente a ellos"* se ha establecido el siguiente objetivo Data Mining*:*

i. Objetivo de Data Mining 3
Para alcanzar este objetivo se construirá un modelo de clasificación según la variable "id_ambito", que se corresponde con el sector al que pertenece la institución. Dicha variable "id_ambito" se obtendrá combinando información de las tablas T_VIRUS, inci_virus e inci_sensores.

1.3.1 Criterios de Éxito de Data Mining

Nº	Objetivo DM	Medida de Éxito
1	Obtener una clasificación para la clase "tipo"	Obtener un modelo de clasificación con un conjunto de entrenamiento formado por el 75 % de los datos de la fuente de origen T_VIRUS y un conjunto de testeo con el 25 % restante, y lograr un porcentaje de éxito mayor del 70% y resultados óptimos en las medidas de precisión[3] y cobertura[4].
2	Obtener una clasificación para la clase "cluster" previamente generada	Obtener un modelo de clasificación con un conjunto de entrenamiento formado por el 75 % de los datos de la fuente de datos T_VIRUS y un conjunto de testeo con el 25 % restante y lograr un porcentaje de éxito mayor del 70% y resultados óptimos en las medidas de precisión y cobertura.
3	Obtener una clasificación para la clase "id_ámbito"	Obtener un modelo de clasificación con un conjunto de entrenamiento formado por el 75 % de los datos de la fuente de datos T_VIRUS2 y un conjunto de testeo con el 25 % restante y lograr un porcentaje de éxito mayor del 70% y resultados óptimos en las medidas de precisión y cobertura.

Tabla 15: Criterios de Éxito de Data Mining

El porcentaje de éxito establecido de 70%, es un criterio orientativo, ya que es probable que no se consiga esta medida de éxito debido a que , como se ha mencionado antes, los objetivos de negocio son ficticios por no disponer de un cliente final.

[3] La precisión mide el número de términos correctamente reconocidos respecto al total de términos predichos.

[4] La cobertura mide la proporción de términos correctamente reconocidos respecto al total de términos reales.

1.4. Plan de Proyecto

Para la planificación del proyecto, primero se ha especificado una lista de tareas a realizar a lo largo de todo el proyecto.

Nº Tarea	Descripción de la tarea
1	Entender el dominio del problema y los datos.
2	Aprender a manejar la herramienta WEKA.
3	Aprender a utilizar y conocer la metodología *"CRISP-DM"*.
4	Plantear el problema.
5	Seleccionar el subconjunto de datos a estudiar.
6	Limpieza, preparación y transformación de los datos.
7	Aplicar la(s) técnica(s) de data mining.
8	Evaluar los resultados obtenidos y refinar el proceso incluyendo posibles iteraciones de los pasos 5 a 8.
9	Generar el documentos final.

Tabla 16: Lista de tareas

Una vez definidas dichas tareas y siguiendo el mismo orden en el desarrollo del proyecto se ha elaborado el correspondiente diagrama de Gantt. La planificación y el tiempo dedicado a cada tarea se han planificado teniendo el cuenta dos de los requisitos del trabajo:

- 324 horas de trabajo en total
- Fecha límite de entrega del trabajo: 9 de Junio de 2013 a las 23:55 h.

Diagrama de Gantt

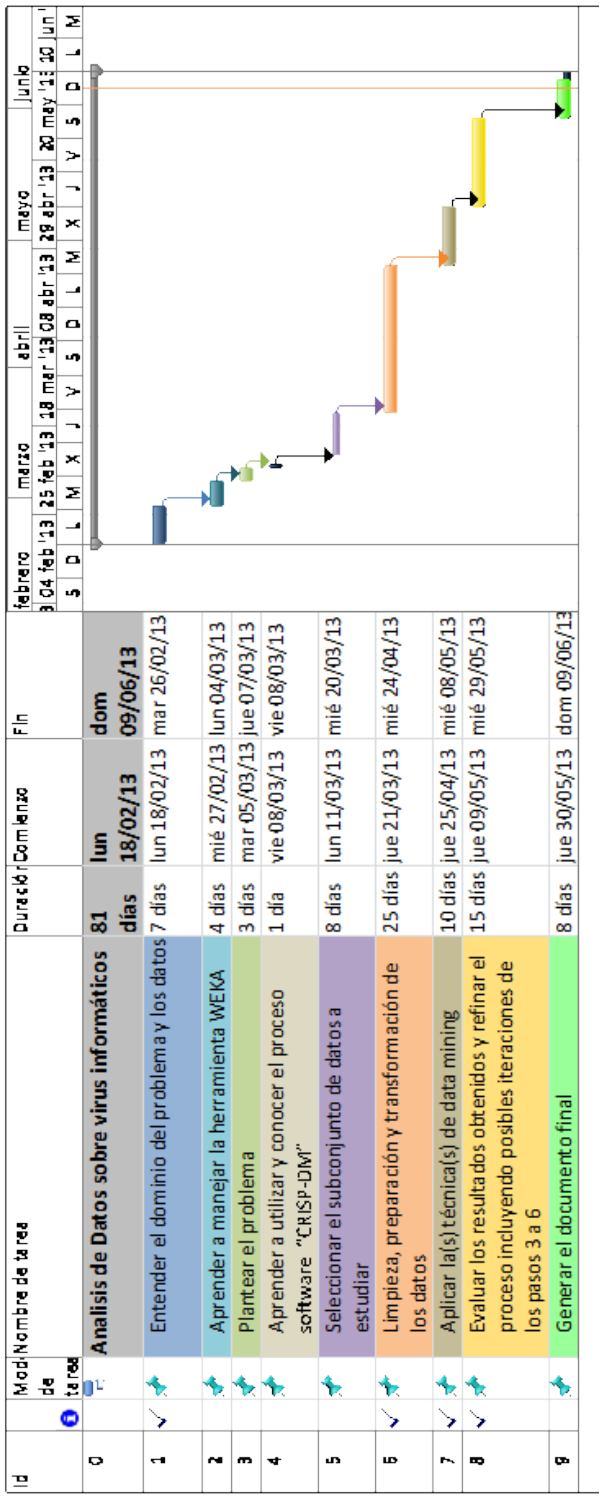

Figura 2: Diagrama de Gantt

Planificación de Tareas

Con el fin de llevar un buen seguimiento de la realización de cada tarea, en la siguiente figura se especifica la duración, comienzo y fin de cada una de las tareas que forman el proyecto. El número total de horas exigidas para la realización del trabajo está repartido cuidadosamente entre cada tarea.

Se puede observar que la tarea que más tiempo ha necesitado es la número 6, la fase correspondiente a la limpieza, preparación y transformación de los datos, como es normal en un proyecto de Data Mining.

Id		Modo de tarea	Nombre de tarea	Trabajo	Duración	Comienzo	Fin
0			**Analisis de Datos sobre virus informáticos**	**324 horas**	**81 días**	**lun 18/02/13**	**dom 09/06/13**
1	✓		Entender el dominio del problema y los datos	24 horas	7 días	lun 18/02/13	mar 26/02/13
2			Aprender a manejar la herramienta WEKA	16 horas	4 días	mié 27/02/13	lun 04/03/13
3			Plantear el problema	15 horas	3 días	ar 05/03/13	jue 07/03/13
4			Aprender a utilizar y conocer el proceso software "CRISP-DM"	4 horas	1 día	vie 08/03/13	vie 08/03/13
5			Seleccionar el subconjunto de datos a estudiar	32 horas	8 días	lun 11/03/13	mié 20/03/13
6	✓		Limpieza, preparación y transformación de los datos	100 horas	25 días	jue 21/03/13	mié 24/04/13
7	✓		Aplicar la(s) técnica(s) de data mining	44 horas	10 días	e 25/04/13	mié 08/05/13
8	✓		Evaluar los resultados obtenidos y refinar el proceso incluyendo posibles iteraciones de los pasos 3 a 6	60 horas	15 días	jue 09/05/13	mié 29/05/13
9			Generar el documento final	29 horas	8 días	e 30/05/13	dom 09/06/13

Figura 3: Tareas a realizar

2. COMPRENSION DE LOS DATOS

La información de origen ha sido recopilada de diferentes empresas anónimas, con el propósito de llevar a cabo un estudio de extracción de conocimiento para predicción de virus informáticos.

2.1 Recogida de los datos iniciales

Inicialmente se dispone de seis fuentes de datos diferentes:

	Nombre	Descripción	Registros	Tamaño
1	T_ALIAS	*Relaciona cada virus con sus diferentes alias posibles*	14.335	674.5 KB
2	T_VIRUS	*Proporciona información sobre cada virus*	6.347	27.0 MB
3	inci_virus	*Recoge el número de incidencias de cada alias(virus).*	1.801.456	152.9 MB
4	inci_informes	*Recoge los datos sobre cada uno de los informes generados tras el análisis diario de virus en el correo electrónico.*	68.371	12.3 MB
5	inci_sensores	*Recoge los sensores en los cuales se analiza el correo electrónico en busca de virus.*	82	24.2 MB
6	inci_ambitos	*Recoge los ámbitos en los cuales se analiza el correo electrónico en busca de virus.*	7	7 MB

Tabla 17: Fuentes de datos

Todas las fuentes de datos proporcionadas han sido presentadas en formato delimitado por comas (csv), pero no todas serán utilizadas en el análisis. Para lograr el objetivo de Data Mining 1 y 2 será necesaria la tabla T_VIRUS, y para lograr el objetivo de Data Mining 3 las tablas T_VIRUS, inci_virus e inci_sensores.

Se describirán los atributos de todas las tablas con el fin de entender mejor la información de la que se dispone. Sin embargo, sólo se detallarán y estudiarán los valores de atributos de las tres tablas (*T_VIRUS*, *inci_virus* e *inci_sensores)*necesarias para cubrir los

objetivos propuestos de data mining. Para las restantes tres tablas que no serán utilizadas, no se describirán los valores de sus atributos. Más adelante se explicará la razón por la que se incluye y excluye cada fuente de datos.

En la *Figura* ***4*** se visualiza el conjunto global de las seis fuentes de información.

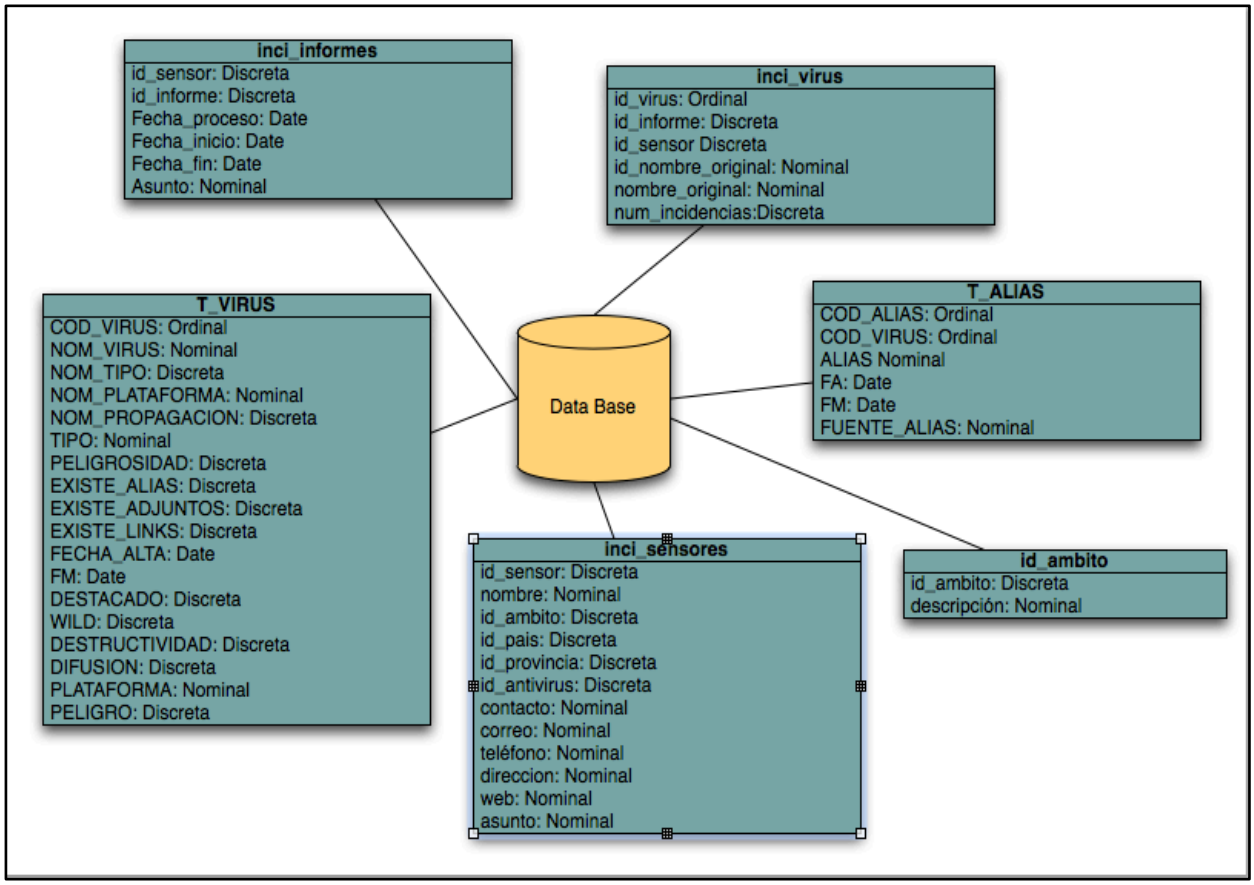

Figura 4:Conjunto de las seis fuentes de datos

2.2 Descripción de los datos

6. Fuente de datos T_ALIAS

	Atributo	Tipo	Descripción
1	COD_ALIAS	Ordinal	Código alias del virus
2	COD_VIRUS	Ordinal	Código del virus (se corresponde con el COD_VIRUS de la tabla T_VIRUS)
3	ALIAS	Nominal	Alias del virus
4	FA	Date	Fecha en la que se dio en alta el alias
5	FM	Date	Fecha en la que se modificó información del alias.
6	FUENTE_ALIAS	Nominal	Fuente de la que proviene el alias

Tabla 18: T_ALIAS

- ***Descripción de los valores de los atributos***

No se detallarán los valores de los atributos de esta tabla debido a que no se utilizará en el análisis.

6. Fuente de datos T_VIRUS

En la siguiente tabla se recogen cada uno de los atributos que forman la tabla T_VIRUS, se muestra su tipo de variable y significado (descripción del atributo).

	Atributo	Tipo	Descripción
1	COD_VIRUS	Ordinal	Código del virus
2	NOM_VIRUS	Nominal	Nombre completo del virus
3	NOM_TIPO	Discreta	Representa con un número el tipo de virus
4	NOM_PLATAFORMA	Discreta	Número de plataforma sobre la que actúa el virus
5	NOM_PROPAGACION	Discreta	Valor de propagación que utiliza el virus para propagarse.
6	TIPO	Nominal	Indica simbólicamente el tipo del virus
7	PELIGROSIDAD	Discreta	Representa con un valor la peligrosidad del virus
8	EXISTS_ALIAS	Boolean	Indica si el virus tiene alias
9	EXISTS_ADJUNTOS	Boolean	Indica si el virus tiene adjuntos
10	EXISTS_LINKS	Boolean	Indica si el virus tiene referencias
11	FECHA_ALTA	Date	Fecha de alta en la que se interceptó el virus
12	FM	Date	Fecha en la que ha se modificado información sobre el virus.
13	DESTACADO	Discreta	Indica si el virus es destacado
14	WILD	Discreta	Indica lo salvaje que es el virus
15	DESCTRUCTIVIDAD	Discreta	Indica la destructividad del virus
16	DIFUSIÓN	Discreta	Indica la difusión del virus
17	PLATAFORMA	Nominal	Nombre de la plataforma sobre la que actúa el virus.
18	PELIGRO	Discreta	Identifica numéricamente el peligro del virus

Tabla 19: Tabla T_VIRUS

- ***Descripción de los valores de los atributos***

A continuación se detallan los valores que toma cada uno de los atributos para los 6347 registros de la tabla T_VIRUS, indicando cuántos de esos registros toman valor *null* para el atributo, cuántos valor 0 y cuántos aparecen en blanco (es decir, no se ha recogido el valor para ese atributo).

	Atributo	Valores	Null	0	Missing
1	COD_VIRUS	De 1 a 6347	0	0	0
2	NOM_VIRUS	6310 valores distintos	0	0	0
3	NOM_TIPO	De 0 a 17	315	56	0
4	NOM_PLATAFORMA	De0a45	2112	166	0
5	NOM_PROPAGACION	De 0 a 18	2121	2566	0
6	TIPO	Virus, VBS, Virus de macro, Gusano, Troyano, Hoax, Otros	0	0	0
7	PELIGROSIDAD	De 1 a 10	5180	0	0
8	EXISTS_ALIAS	Todos null	6347	0	0
9	EXISTS_ADJUNTOS	Todos null	6347	0	0
10	EXISTS_LINKS	Todos null	6347	0	0
11	FECHA_ALTA	6323 valores únicos	0	0	0
12	FM	4685 valores distintos	1650	0	0
13	WILD	Bajo, Medio, Alto	2002	4254	1
14	DESCTRUCTIVIDAD	Bajo, Medio, Alto	1901	150	4
15	DIFUSIÓN	Bajo, Medio, Alto	1901	146	4
16	PLATAFORMA	231 valores distintos	1757	0	2878
17	PELIGRO	De 0 a4	0	0	0

Tabla 20: Descripción de atributos de T_VIRUS

6. Fuente de datosinci_virus

	Atributo	Tipo	Descripción
1	**Id_virus**	Ordinal	Identificador del virus (se corresponde con el id_virus de la tabla T_VIRUS)
2	**Id_informe**	Discreta	Identificador del informe (se corresponde con el id_informe de la tabla inci_informes)
3	**Id_sensor**	Discreta	Identificador del sensor (correspondiente con el id_sensor de la tabla inci_sensores)
4	**Id_nombre_original**	Nominal	Identificador del nombre original del virus
5	**Nombre_original**	Nominal	Nombre original del virus
6	**Num_incidencias**	Discreta	Número de incidencias registradas del virus

Tabla 21: Fuente de datos inci_virus

- ***Descripción de los valores de los atributos***

	Atributo	Valores	Null	0	Missing
1	**Id_virus**	919 valores distintos	0	0	0
2	**Id_informe**	De 1 a 3074	0	0	0
3	**Id_sensor**	De 1 a 106	0	0	0
4	**Id_nombre_original**	De 1 a 7	0	0	0
5	**Nombre_original**	De 1 a 667	0	0	0
6	**Num_incidencias**	15076 valores distintos	0	0	0

Tabla 22 : Descripción de atributos de inci_virus

6. Fuente de datos inci_informes

	Atributo	Tipo	Descripción
1	Id_sensor	Discreta	Identificador del sensor (se corresponde con el id de la tabla inci_sensores)
2	Id_informe	Discreta	Identificador del informe
3	Fecha_proceso	Date	Fecha del proceso
5	Fecha_inicio	Date	Fecha inicio del proceso
6	Fecha_fin	Date	Fecha fin del proceso
7	Asunto	Nominal	Asunto del informe

Tabla 23: Fuente de datos inci_informes

- ***Descripción de los valores de los atributos***

No se detallarán los valores de los atributos de esta tabla debido a que este origen de información no se utilizará en el análisis.

6. Fuente de datos inci_sensores

	Atributo	Tipo	Descripción
1	**id_sensor**	Discreta	Identificador de la institución (se corresponde con el id_ambito de la tabla inci_ambitos)
2	**nombre**	Nominal	Nombre del sensor(se corresponde a una institución)
3	**id_ambito**	Discreta	Identificador del ámbito
4	**id_pais**	Discreta	Identificador del país
5	**id_region**	Discreta	Identificador de la región
6	**id_provincia**	Discreta	Identificador de la provincia
7	**id_antivirus**	Discreta	Identificador del antivirus
8	**contacto**	Nominal	Persona de contacto para esta institución
9	**correo**	Nominal	Correo electrónico de la persona de contacto
10	**dirección**	Nominal	Dirección de la institución
11	**web**	Nominal	Pagina web de la institución

Tabla 24:Fuente de datos inci_sensores

- ***Descripción de los valores de los atributos***

Sólo se detallarán dos de los atributos de esta tabla, ya que únicamente ellos dos serán utilizados en el análisis de datos.

	Atributo	Valores	Null	0	Missing
1	**Id_sensor**	De 1 a 106	0	0	0
2	**Id_ambito**	De 1 a 7	0	0	0

Tabla 25: Descripción de atributos de inci_sensores

6. Fuente de datos inci_ambitos

	Atributo	Tipo	Descripción
1	**id_ambito**	Discreta	Identificador del ámbito
2	**descripción**	Nominal	Descripción del ámbito

Tabla 26: inci_ambitos

- ***Descripción de los valores de los atributos***

La tabla inci_ambitos tiene únicamente los siguientes valores:

Id_ambito	Descripción
1	Universidad
2	Administración Central
3	Administración Autonómica
4	Administración Local
5	Sector Público
6	Proveedores de acceso a Internet
7	Internacionales

Tabla 27: Descripción de atributos de inci_ambitos

2.3 Informe exploración de datos

La exploración de los datos se realizará solamente sobre la fuente de datos T_VIRUS, ya que los atributos necesarios de las restantes dos tablas (se utilizarán dos atributos de cada una de estas fuentes que serán necesarios para cubrir el objetivo de data mining 3)no presentan anomalías, ningún valor *null* ni *missing*, tal como se puede observar en la *Tabla **22** : Descripción de atributos de inci_virus* y la *Tabla **25**: Descripción de atributos de inci_sensores*.

2.3.1 Fuente de datos T_VIRUS

i. Tal como se puede observar en la *Tabla **20**: Descripción de atributos de T_VIRUS* existen varios atributos sólo con valores *null*, valores que no han sido especificados por las diferentes empresas.

Nº	Atributo	Valores
9	EXISTE_ALIAS	Todos los registros son *null*
10	EXISTS_ADJUNTOS	Todos los registros son *null*
11	EXISTS_LINKS	Todos los registros son *null*

Tabla 28: Atributos sólo con valor null

ii. No existen registros duplicados puesto que la tabla T_VIRUS proporciona información sobre cada virus, es decir una entrada por cada virus.

iii. Hay atributos con significados parecidos :

a) NOM_TIPO y TIPO
El atributo NOM_TIPO refleja el tipo de virus con un valor numérico, mientras que el atributo TIPO es nominal.
Se ha comprobado que no existe ninguna correlación entre ellos (los registros con el atributo NOM_TIPO 1 muestran tipo de virus Gusano, Troyano, Otros), por lo tanto, son atributos que reflejan cosas diferentes. Además, NOM_TIPO tiene 315 registros con valores *null* y el atributo TIPO tiene todos sus registros con valores válidos.

b) PELIGROSIDAD y PELIGRO.
El atributo PELIGROSIDAD tiene valores de 1 al 9 y el atributo PELIGRO refleja valores del 1 al 4.
Se ha comprobado que no existe ninguna correlación entre ellos (registros con PELIGROSIDAD 2 tiene valores de PELIGRO 1 ,2 y 4), por lo tanto son atributos que reflejan cosas diferentes. Además, el atributo PELIGROSIDAD posee 5180

registros con valor *null* y el atributo PELIGRO tiene todos sus registros con valores válidos.

c) NOM_PLATAFORMA y PLATAFORMA
El atributo NOM_PLATAFORMA tiene valores de 0 a 47 y el atributo PLATAFORMA es de tipo nominal (ej: Windows 98,Linux). Mientras que NOM_PLATAFORMA no muestra ningún valor perdido, el atributo PLATAFORMA tiene 1757 registros con valores *null* y 2878 con valores perdidos.
Registros que tienen NOM_PLATAFORMA = 10 muestran 12 distintas PLATAFORMAS, por lo tanto, son atributos que no presentan correlación entre ellos.

iv. Dada la gran cantidad de atributos de la tabla es posible que haya algunas variables que estén correlacionadas con otras. *Weka* proporciona el modo *visualización* , ver *Figura* ***5***, que presenta gráficamente la distribución de todos los atributos mostrando gráficas en dos dimensiones, en las que en los ejes se va representando todos los posibles pares de combinaciones de atributos. Este modo permite fácilmente detectar correlaciones y asociaciones entre los atributos.

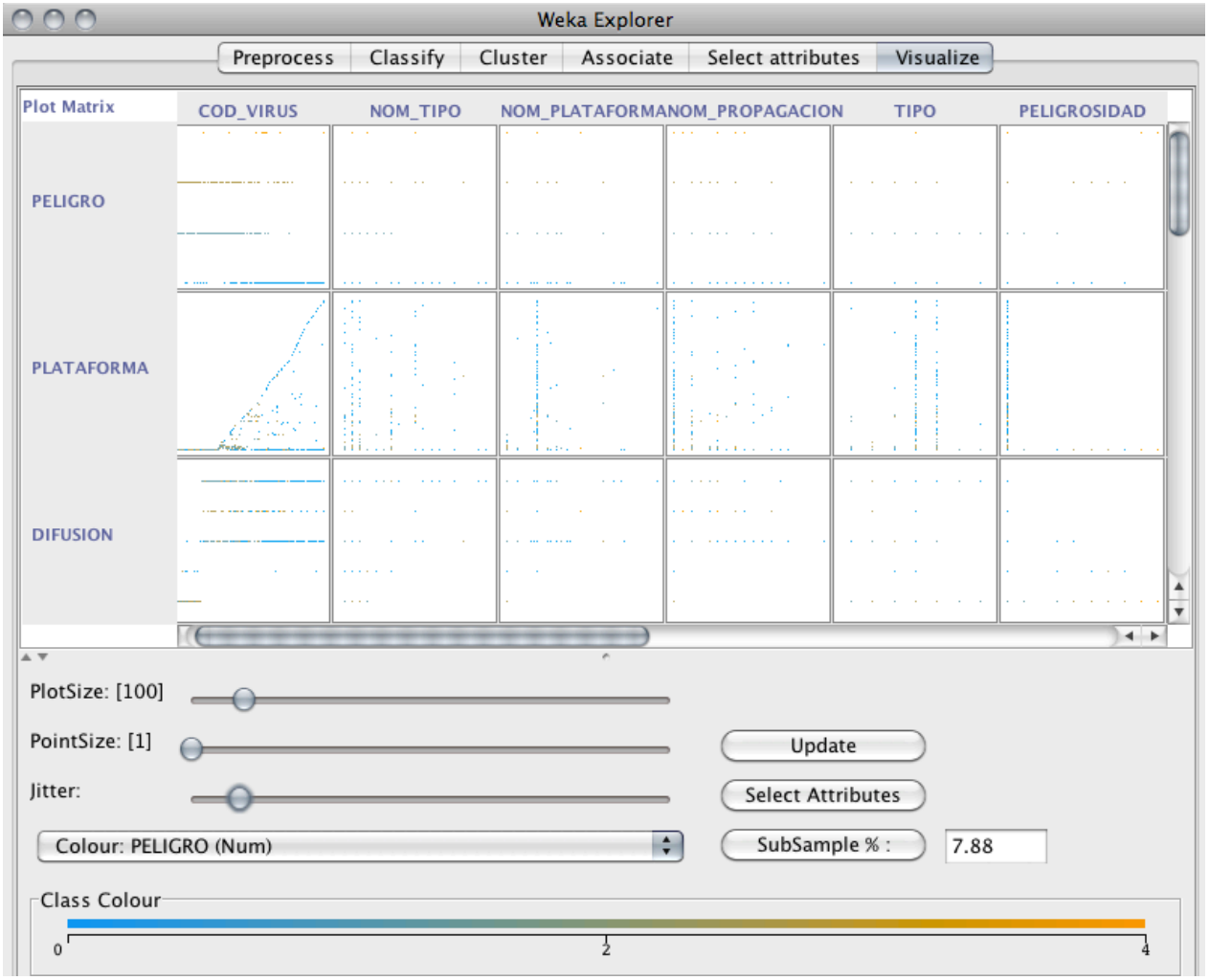

Figura 5: Modo Visualización WEKA

Sin embargo, debido al gran desbalance de los datos, muchos registros nulos y otros con valor 0, no se aprecia ninguna correlación significativa entre los atributos.

v. Con la opción de visualización que proporciona WEKA se pueden observar algunos datos importantes sin tener que utilizar técnicas más complejas para ello.

En la *Figura* ***6*** se observa que los gusanos (color turquesa) son los virus que mayor nivel de difusión y destructividad muestran. Dichos factores son importantes a la hora de establecer lo dañino que es el virus. Es decir, que se puede sostener que los gusanos son el tipo de virus más dañino de los sietes tipos debido a su alta difusión y destructividad.

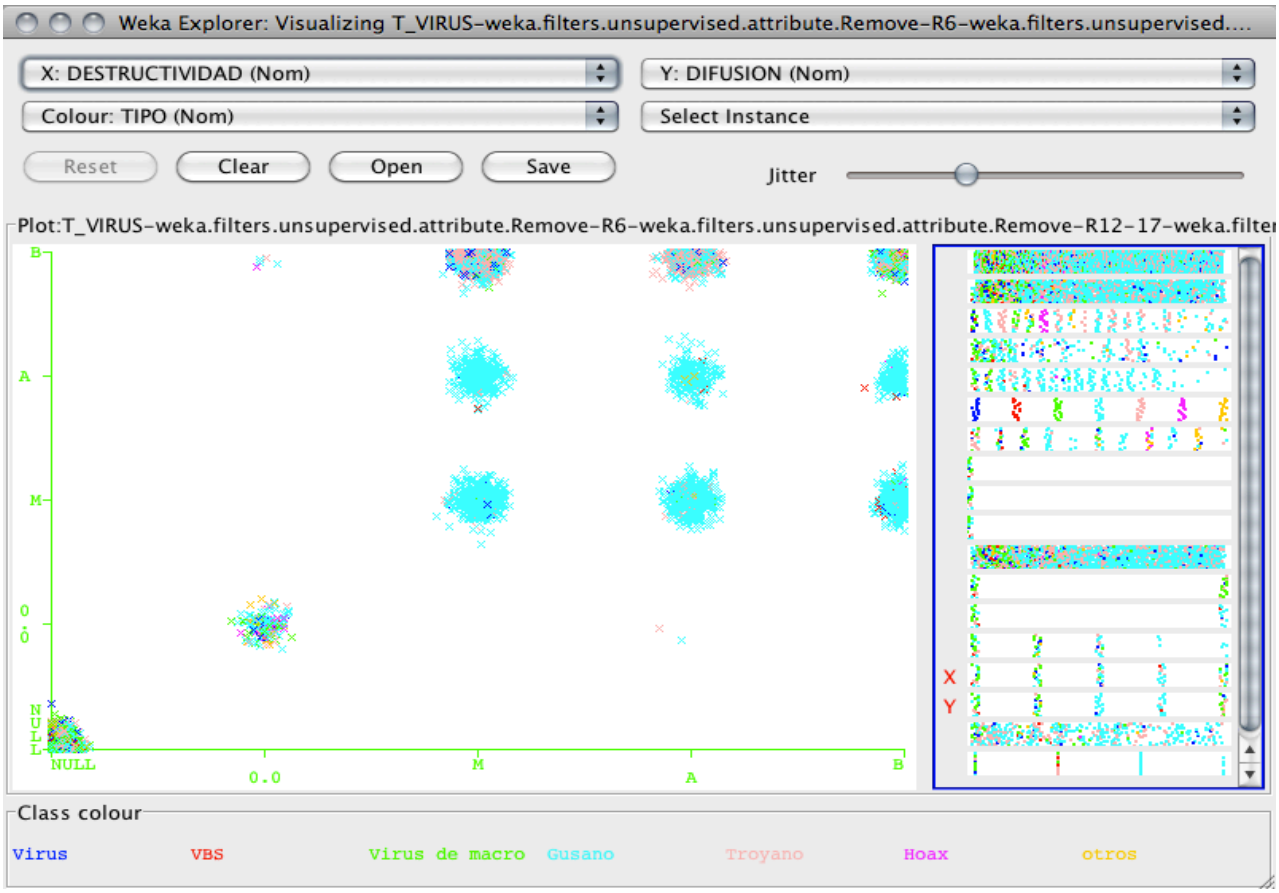

Figura 6: Visualización 2D-1

En la siguiente ilustración, se observa que los Gusanos (turquesa) son los únicos virus que poseen un peligro (eje X) con valor 4. Con el valor 3 hay presencia de virus VBS (rojo), Virus de macro (verde) y Troyano (rosa), pero el Gusano sigue siendo el virus predominante con diferencia.

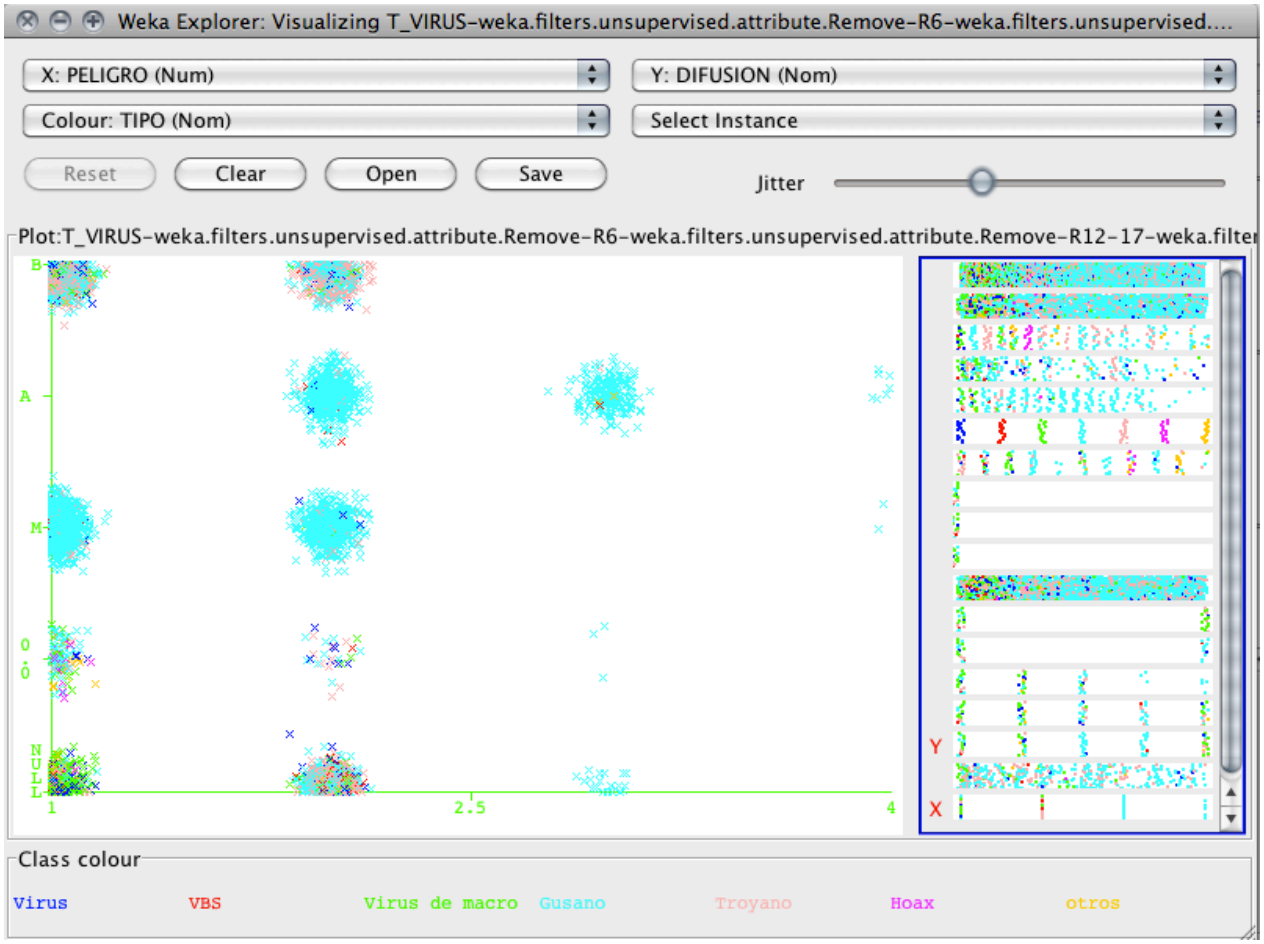

Figura 7: Visualización 2D-2

Por último, en la *Figura **8**: Visualización 2D-3*, se contempla en el eje Y que hay pocos virus que tienen el valor Sí como destacado (el atributo Destacado tiene únicamente dos valores , Sí o No), mostrando también en el eje X su valor de peligro. Como era de esperar, el gusano es un virus que destaca y también el único que tiene peligro 3. Con peligro 2 y 1 se aprecian los mismos virus que se describieron en la ilustración anterior: VBS (rojo), Virus de macro (verde) y Troyano (rosa)con menor número que los gusanos.

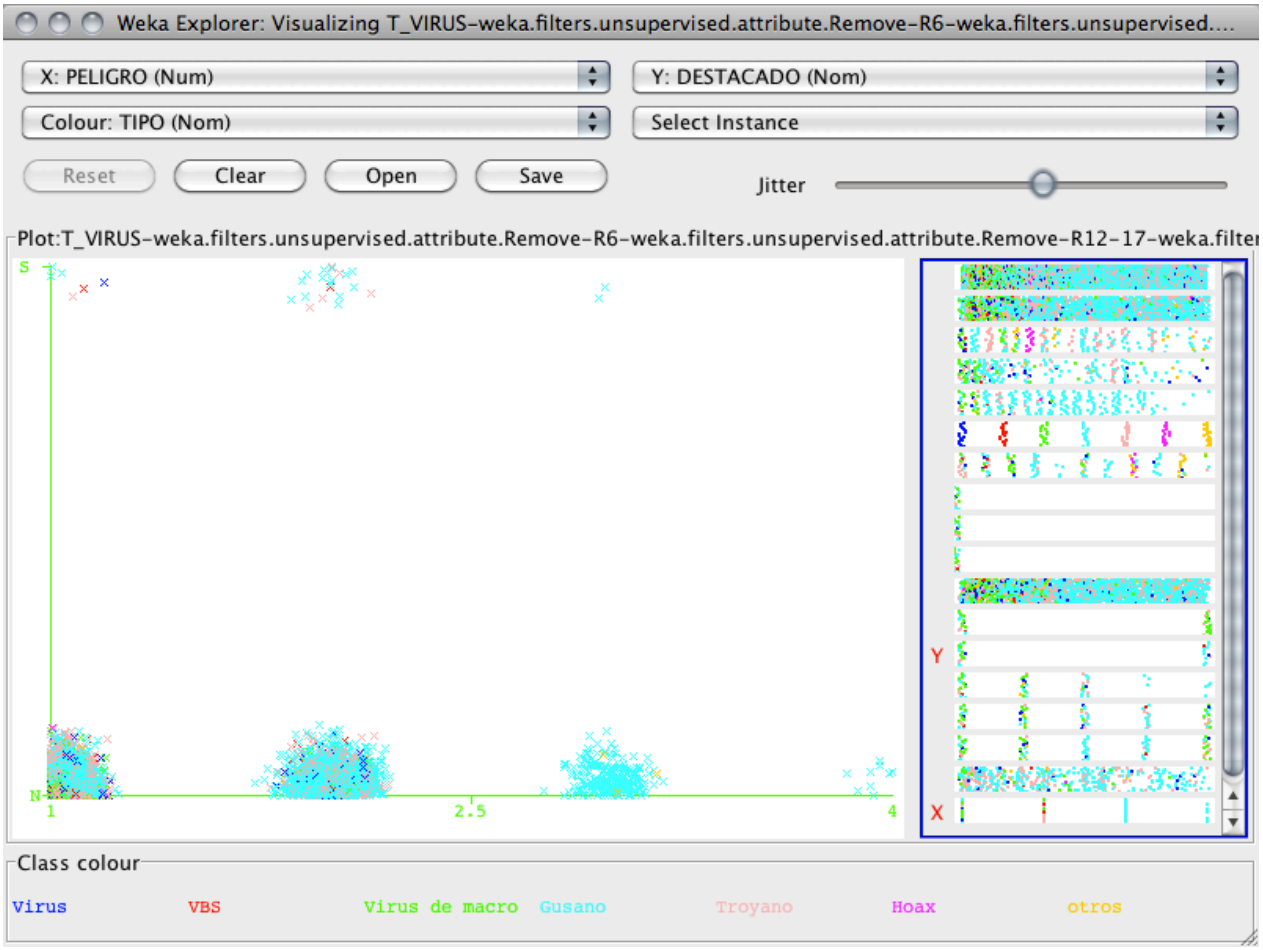

Figura 8: Visualización 2D-3

2.4 Verificación de la Calidad de los Datos

Conforme con los datos expuestos en las tablas de descripción de los datos, se pueden sacar las siguientes conclusiones.

Calidad	
Errores tipográficos	No se observan errores tipográficos
Atributos redundantes	No se observan atributos redundantes
Atributos que no poseen ningún valor *null.*	NOM_VIRUS TIPO FECHA_ALTA PELIGRO
Atributos sólo con valores *null*	EXISTS_ALIAS EXISTS_ADJUNTOS EXISTS_LINKS
Atributos que poseen algún valor *null* (en porcentaje)	NOM_TIPO 4,96% NOM_PLATAFORMA 33,5% NOM_PROPAGACION 33,5% PELIGROSIDAD 82% FM 26% WILD 32% DESTRUCTIVIDAD 30% DIFUSIÓN 30% PLATAFORMA 28%
Atributos que contienen valores *"missing" o en blanco* (en porcentaje)	WILD 0,01% DESTRUCTIVIDAD 0,06% DIFUSION 0,06% PLATAFORMA 28%
Atributos nominales que no deberían tener valor 0	Los atributos WILD, DESTRUCTIVIDAD y DIFUSION tienen los niveles bajo(B), medio(M) o alto(A), por lo tanto , habiendo también valores *null* , el valor 0 se entendería como muy bajo o desconocido.

Tabla 29: Informe de calidad de datos

3. PREPARACIÓN DE LOS DATOS

3.1 Selección de los Datos

3.1.1 Selección de fuentes de datos

Inicialmente se dispone de seis fuentes de datos, seis tablas diferentes , pero sólo se utilizarán tres de ellas debido a quelas otras tres no contienen información útil para el proceso de minería de datos y no pueden cubrir los objetivos establecidos de Data Mining.

id	Tabla	*Incluida*	*Razón*
1	T_ALIAS	No	*No se utilizarán los alias de virus en el análisis.*
2	T_VIRUS	*Sí*	*Proporciona información relevante y primordial sobre cada virus detectado.*
3	inci_virus	*Sí*	*En esta tabla se refleja el número de incidencias de virus que han sufrido las diferentes instituciones y empresas.*
4	inci_informes	*No*	*Esta fuente de datos contiene información sobre cada uno de los informes generados por incidencia de virus, El contenido de los informes no aportan información relevantes para el análisis de datos.*
5	inci_sensores	*Sí*	*Se utilizará la información correspondiente a id_sensor e id_ambito. De esta manera se podrá añadir el campo id_ambito a la tabla inci_informes, que dispone de la información relativa al sensor pero no del ámbito del sensor.*
6	inci_ambitos	No	*Esta tabla recoge los ámbitos de todas las organizaciones y empresas, distinguiendo entre siete diferentes.*

Tabla 30: Selección de datos

3.1.2 Selección de atributos de cada fuente de datos seleccionada

- ***Fuente de datos T_VIRUS***

	Atributo	Incluido	Razón
1	COD_VIRUS	Si	Atributo con el código de virus (1 -6347). El código de virus se usará para posteriormente enlazar la tabla *T_VIRUS* con la fuente de datos *inci_virus*.
2	NOM_VIRUS	No	Atributo con el nombre del virus. No es relevante para el análisis de datos puesto que son 6347 nombre diferentes.
3	NOM_TIPO	No	Identifica el tipo de virus con un número (0-17 y 100). Para conocer el tipo de virus de utilizará el atributo *TIPO*.
4	NOM_PLATAFORMA	Si	Identifica la plataforma sobre la que actúa el virus con un numero de 0 – 45. Hay 2112 valores *null*.
5	NOM_PROPAGACION	Si	Identifica el modo de propagación del virus con un numero de 0 – 18. Hay 2121 valores *null*.
6	TIPO	Si	Atributo con el tipo de virus (Virus, VBS, Virus de macro, Gusano Troyano, Hoax y Otros).
7	PELIGROSIDAD	No	El 80% de los registros son *null*. Atributo excluido de la selección de datos.
8	EXISTE_ALIAS	No	Atributo con todos los registros null, por lo que no proporciona ninguna información. Atributo excluido de la selección de datos.
9	EXISTS_ADJUNTOS	No	Atributo con todos los registros *null*, por lo que no aporta ninguna información. Atributo excluido de la selección de datos.
10	EXISTS_LINKS	No	Atributo con todos los registros *null*, por lo que no aporta ninguna información. Atributo excluido de la selección de datos.
11	FECHA_ALTA	No	Fecha en la que se interceptó el virus. No es una característica relevante para el análisis.
12	FM	No	Casi todas las fechas en la que se realizó una modificación de la información del virus son diferentes, por lo que este atributo no es relevante.

	Atributo	Incluido	Razón
13	DESTACADO	No	Indica si el virus es destacado o no. Esta información no tiene mucha relevancia ya que sólo 29 de los 6347 registros tienen valor *No*, es decir sólo 29 son no destacados. Atributo excluido de la selección de datos.
14	WILD	Si	Indica lo salvaje que es el virus. Hay 2002 registros *null* y 1 *missing*.
15	DESCTRUCTIVIDAD	Si	Refleja la destructividad del virus {Baja, Media, Alta}.
16	DIFUSION	Si	Refleja la difusión del virus {Baja, Media, Alta}.
17	PLATAFORMA	No	Nombre de la plataforma en la actuó el virus. Ej: Windows NT, Linux... Sin embargo, la información que proporciona este atributo no es muy fiable, puesto que hay 2878 valores *missing* y 1757 *null*, en total 4653 registros que no son útiles. Atributo excluido de la selección de datos.
18	PELIGRO	Si	Identifica de 1-4 el peligro del virus.

Tabla 31: Selección de atributos de T_VIRUS

Es importante destacar que, puesto que no contamos con el apoyo de la empresa, no se dispone de ninguna información que pueda mostrar la importancia real de los atributos seleccionados, ni tampoco para la exclusión de los restantes. Dicha inclusión y exclusión ha sido realizada siguiendo el *Informe de calidad de datos*, considerando la calidad de los datos al nivel requerido por las técnicas de análisis seleccionadas.

- ***Fuente de datos Inci_virus***

	Atributo	Incluido	Razón
1	Id_virus	Si	El identificador del virus se corresponde con el cod_virus de la tabla T_VIRUS. De esta forma las dos tablas se pueden enlazar por esta clave.
2	Id_informe	No	El identificador del informe no se utilizará.
3	Id_sensor	Si	El identificador del sensor se utilizará para poder enlazar el sensor con su correspondiente ámbito de la tabla inci_ambitos.
4	Id_nombre_original	No	Identificador del nombre original del virus. Dicho identificador sólo se puede encontrar en esta tabla, por lo que no se puede enlazar con ningún atributo de la tabla T_VIRUS y por consiguiente no es relevante.
5	Nombre_original	No	El nombre original del virus es información adicional. Dicha información también se puede encontrar en la tabla T_VIRUS, concretamente en el atributo nom_virus.
6	Num_incidencias	Si	Es el número de incidencias registradas del virus.

Tabla 32: Selección de atributos de inci_virus

- ***Fuente de datos Inci_sensores***

De esta fuente únicamente se utilizarán dos atributos, id_sensor e id_ambito ya que estos dos atributos se utilizarán para combinar la tabla inci_sensores con la tabla inci_virus.

	Atributo	Incluida	Razón
1	**id_sensor**	Si	Identificador del sensor (se corresponde con el id_sensor de la tabla inci_ambitos)
2	**id_ambito**	Si	Identificador del ámbito. Este atributo será añadido a la tabla inci_virus primero y más tarde a la tabla T_VIRUS

Tabla 33: Selección de atributos de inci_sensores

3.1.2 Selección final de fuentes de datos

Tras haber estudiado los criterios de inclusión/exclusión de cada una de las fuentes de datos y también de su información, los datos que se van a utilizar se muestran en la siguiente ilustración.

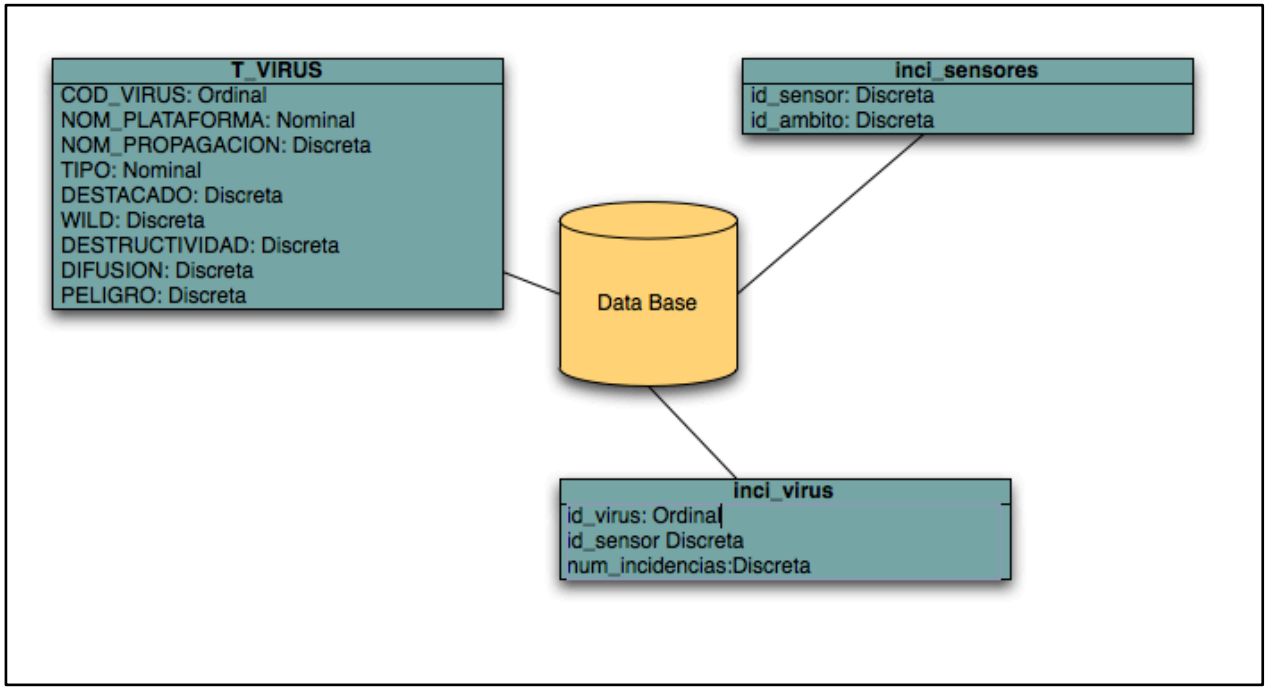

Figura 9: Selección de datos final

3.2 Limpieza de los Datos

En el apartado 3.1 Selección de los Datos se han especificado las fuentes de datos seleccionadas. Así mismo en el apartado 3.1.2 Selección de atributos de cada fuente de datos seleccionada se ha desarrollado la correspondiente selección de atributos a cada fuente de datos, indicando si los atributos de la fuente han sido incluidos o excluidos de la selección de datos. En la siguiente tabla se puede ver de manera resumida cuántos atributos han sido excluidos de cada fuente de datos:

Fuente de datos T_VIRUS

id	Tabla	*Total atributos*	*Atributos incluidos*	*Atributos excluidos*
2	**T_VIRUS**	18	11	7

Tabla 34: Selección de atributos de T_VIRUS

id	Tabla	*Total instancias*	*Instancias eliminadas*	*Instancias utilizadas*
2	**T_VIRUS**	6347	0	6347

Tabla 35: Selección de instancias de T_VIRUS

No se han eliminado instancias de la fuente de datos T_VIRUS, por lo tanto el número de instancias utilizadas es el total.

Fuente de datos inci_virus

id	Tabla	*Total atributos*	*Atributos incluidos*	*Atributos excluidos*
3	**inci_virus**	6	3	3

Tabla 36: Selección de atributos de inci_virus

id	Tabla	*Total instancias*	*Instancias eliminadas*	*Instancias utilizadas*
3	**inci_virus**	1.801.456	360.604	1.440.852

Tabla 37: Selección de instancias de inci_virus

Se han eliminado algunos registros que no son válidos en esta tabla, porque sus variables están fuera de los rangos permitidos.

i. Registros que poseen un *id_virus* menor que 1 (301.576 rgistros)

ii. Registros con *id_sensor* menor que 1 o igual a null (20 registros)

iii. Registros con un *num_incidencias* menor o igual a 0. (8 registros)

En total se han eliminado 360.604 registros , quedando en total 1.440.852 instancias con valores validos

Fuente de datos inci_sensores

id	Tabla	*Total atributos*	*Atributos excluidos*	*Atributos incluidos*
5	**inci_sensores**	14	12	2

Tabla 38: Selección de atributos de inci_sensores

id	Tabla	*Total instancias*	*Instancias eliminadas*	*Instancias utilizadas*
3	**inci_sensores**	88	0	88

Tabla 39: Selección de instancias de inci_sensores

No se han eliminado instancias de la fuente de datos inci_sensores, por lo tanto el número de instancias utilizadas es el total.

3.3 Construcción de los datos

Esta tarea incluye todas las operaciones de preparación de datos realizadas sobre los atributos existentes y sobre sus valores.

Fuente de datos T_VIRUS

i. No ha sido necesario derivar ninguno de los atributos.

ii. No se han incluido nuevos registros.

iii. Se ha cambiado el nombre de la variable "WILD" por "Salvaje"

Hay tres operaciones que se realizan sobre valores de atributos:

- Transformación de valores *null*
- Transformación de valores 0
- Transformación de valores *missing*

Concretamente y para cada atributo de la tabla T_VIRUS que ha sufrido alguna transformación, las operaciones son las siguientes:

i. Los valores *null* de la variable NOM_PLATAFORMA han sido transformados a valor *"0"*,conservando así la propiedad de atributo numérico.

ii. Los valores *null* de la variable NOM_PROPAGACION han sido transformados a valor *"0"*, conservando así la propiedad de atributo numérico.

iii. Los valores *null* y 0 de la variable WILD han sido transformados con el valor *"Desconocido"*.

iv. Los valores *missing* de la variable WILD han sido transformados con valor de la moda. Para ello se ha utilizado el filtro **ReplaceMissingValues** de WEKA**,** que reemplaza todos los valores indefinidos por la moda, en el caso de un atributo de tipo nominal.

v. Los valores *null* y valores 0 de la variable DESTRUCTIVIDAD han sido transformados con valor *"Desconocido"*.

vi. Los valores *missing* de la variable DESTRUCTIVIDAD han sido transformados con valor de la moda Para ello se ha utilizado el filtro **ReplaceMissingValues** de WEKA**,** que reemplaza todos los valores indefinidos por la moda, en el caso de un atributo de tipo nominal.

vii. Los valores *null* y valores 0 de la variable DIFUSION han sido transformados con el valor *"Desconocido"*.

viii. Los valores *missing* de la variable DIFUSIÓN han sido transformados con el filtro **ReplaceMissingValues**de WEKA**,** que reemplaza todos los valores indefinidos por la moda ya que es un atributo nominal.

Para hacer cada una de estas transformaciones se ha basado en diferentes factores que se han observado para los valores de cada atributo:

Transformación	***Razón***
Transformación de valores *null*	La tabla T_VIRUS contiene 6347 registros de virus, si los valores null de cada atributo se borrasen, se perdería más del 50% de los registros, una cantidad desmesurada de información. Por lo tanto, con el fin no perder esta información, las casillas (valores en columnas) cuyo valor es *null*, se han sustituido por el valor Desconocido. Además, el valor *Desconocido* se puede interpretar claramente como ausencia de valor.
Transformación de valores 0	Los valores 0, en este caso, han sido transformados en atributos simbólicos Salvaje, Destructividad y Difusión que reflejan un nivel (Bajo, Medio y Alto), siendo este valor incorrecto en esta categoría. Además, este valor no tiene interpretación *null* puesto que los valores *null* ya han sido transformados como desconocidos. Por lo tanto, por la misma razón de evitar perder registros, el valor 0 se ha sustituido por el nivel de muy bajo.
Transformación de valores *missing*	Para transformar los valores *missing* se ha empleado e lfiltro *ReplaceMissingValues* que proporciona la herramienta de data mining en el panel de pre-procesamiento, siendo éste el más apropiado. El filtro reemplaza los valores indefinidos del atributo simbólico (sólo los atributos simbólicos finalmente seleccionados Wild, Destructividad y Difusión tienen registros indefinidos) por la moda, el valor con mayor frecuencia en la distribución de datos.

Tabla 40: Transformación de valores

Fuente de datos inci_virus

No se ha realizado ninguna operación de preparación de datos.

Fuente de datos inci_sensores

No se ha realizado ninguna operación de preparación de datos.

3.4 Integración de los datos

Se combinará parte de la información de las tablas T_VIRUS, inci_virus e inci_sensores para crear un nuevo atributo en la tabla T_VIRUS. La información, atributos, que se utilizan de cada una de ellas se ha explicado en el paso anterior *3.1.2 Selección de atributos de cada fuente de datos seleccionada*. El propósito de esta combinación es conseguir obtener el atributo id_ambito(que identifica el ámbito de la institución sobre la que actúa el virus) incorporado en la fuente de datos T_VIRUS. De esta manera, la tabla T_VIRUS contendrá una columna más, que identificará el ámbito (sector en que se agrupan las instituciones) al cual el virus ha atacado. Este atributo será la variable clase en el modelo de clasificación definido en el objetivo de Data Mining 3.

En total se han efectuado dos combinaciones entre pares de tablas, utilizando el paquete XAMPP[5].

3.4.1 Combinación 1: inci_sensores – inci_virus

Observando la información de la *Tabla **27**: Descripción de atributos deinci_ambitos* se aprecia claramente que existen 7 diferentes ámbitos en los que se han interceptado los virus: Universidad, Administración local, Administración Autonómica, Administración Local, Sector Público, Proveedores de acceso a Internet e Internacionales.

Se ha filtrado dicha tabla según el número de id_ambito para conocer exactamente el id_sensor que lo forma. Por ejemplo:

id_sensor	Nombre	id_ambito
94	[redacted][6]	7
95	[redacted]	7

Tabla 41: id_sensor correspondiente a id_ambito 7

El dato correspondiente al identificador del sensor sí está en la tabla de incidencias de virus (inci_virus) pero no el dato del identificador del ámbito.

Se ha añadido un atributo (columna) id_ambito a la tabla inci_virus y se ha actualizado su valor correspondiente para cada conjunto de id_sensor . El *update* se ha ejecutado para cada uno de los siete valores de id_ambito. En el ejemplo de abajo se muestra para el valor 7.

[5] Es un software que integra el servidor Apache, la base de datos MYSQL y PhPMyAdmin.

[6] Los nombres de las instituciones se han suprimidos por motivos de privacidad

```
updateinci_virus
setid_ambito='7'
whereid_sensorin(94,95)
```

3.4.2 Combinación 2: T_VIRUS – ámbito_virus

Lo que se pretende hacer con esta segunda combinación es añadir a la tabla T_VIRUS el ámbito en el cual ha sido interceptado el virus. Para ello, primero se ha creado la tabla ámbito_virus que agrupa todas las incidencias del virus según el ámbito, reduciendo el número de 1.440.852 a 2.601 registros. Dicha tabla tiene únicamente tres columnas, id_virus (se corresponde con el cod_virus de la tabla T_VIRUS), id_ambito e num_incidencias. El atributo num_incidencias indica la suma de incidencias que el virus ha causado en ese ámbito, pero este dato, a pesar de haberse obtenido, no es relevante para el estudio.

Después de este paso , se han combinado la nueva tabla ámbito_virus y la tabla T_VIRUS en una nueva, T_VIRUS2, con *inner join* entre la columna cod_virus de la tabla T_VIRUS y la columna id_virus de ámbito_virus. Finalmente, esta nueva tabla contiene 2594 registros, debido a que únicamente contiene los virus que han creado alguna incidencia en un ámbito.

```
insert into ambito_virus(id_virus,id_ambito, num_incidencias)
select id_virus,id_ambito, sum(num_incidencas) as num_incidencias
from inci_virus
group by id_virus,id_ambito
```

```
Insert into T_VIRUS2
(cod_virus,nom_tipo,nom_plataforma,nom_propagacion,tipo,peligrosidad,destacado,wild,
destructividad, difusion ,peligro, id_ambito)
Select
tv.cod_virus,tv.nom_tipo,tv.nom_plataforma,tv.nom_propagacion,tv.tipo,tv.peligrosidad,
tv.destacado, tv.wild,tv.destructividad,tv.difusion,tv.peligro,ini.id_ambito
from T_VIRUS tv
inner join inci_virusini
on ini.id_virus = tv .cod_virus
```

Las dos combinaciones y pasos seguidos para su construcción se muestran en la siguiente figura.

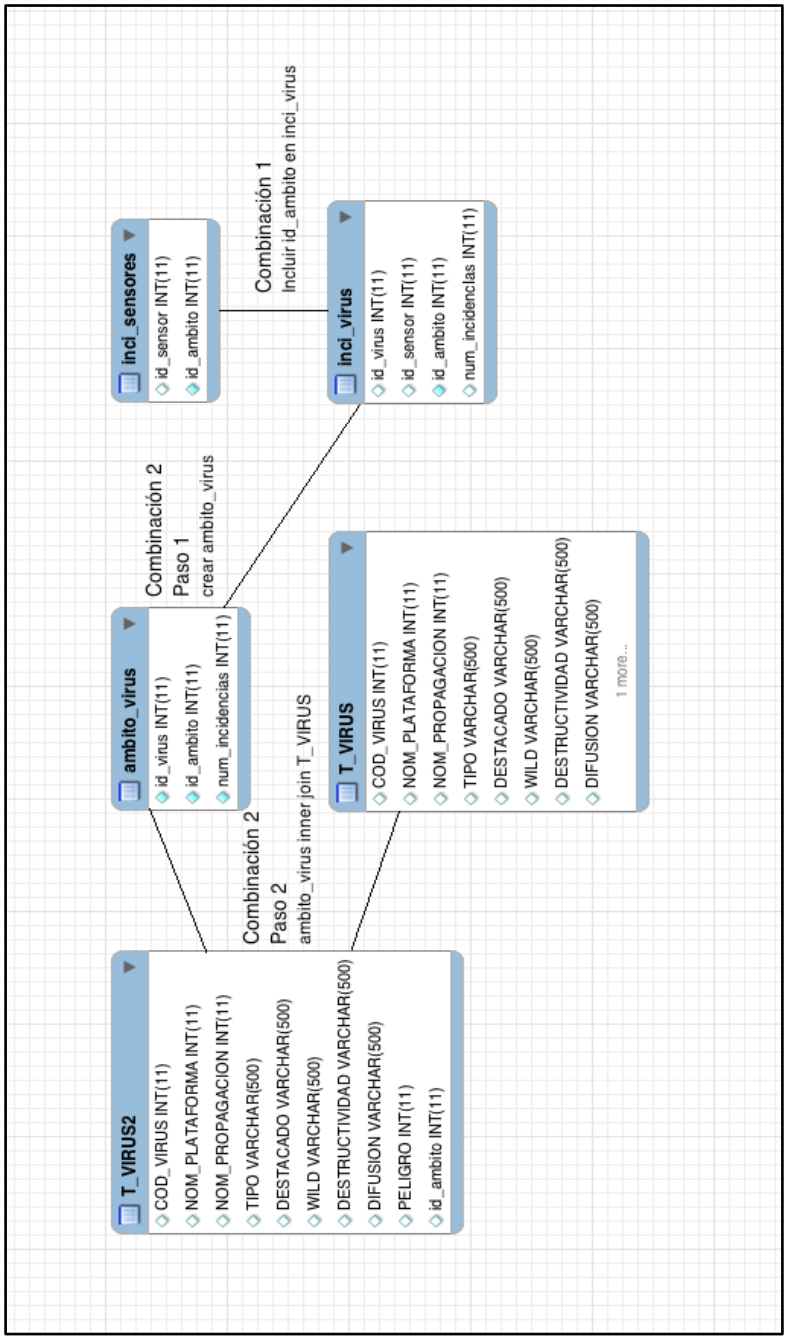

Figura 10: Combinación de datos

3.5 Formateo de los datos

No se han realizado tareas de formateo de datos sobre ninguna de las fuentes de datos.

Además, la herramienta WEKA no tiene ninguna preferencia ni requerimiento sobre el orden de los atributos, por lo que el orden que siguen no tiene relevancia para el procesamiento de los datos.

4. MODELADO

4.1 Selección de las Técnicas de Modelado

En cada objetivo de Data Mining se han seleccionado una o varias técnicas y algoritmos para su desarrollo. Para dejar constancias del algoritmo y técnica específicos utilizados en cada uno de ellos se muestran las siguientes tablas.

4.1.1 Objetivo DM 1

En la siguiente tabla se muestra la técnica utilizada para el objetivo de Data Mining 1 y también el algoritmo elegido para aplicar dicha técnica.

Nº	Objetivo DM	Técnica	Algoritmo
1	Obtener una clasificación para la clase "tipo de virus"	Clasificación	J48 (C4.5)[7]

Tabla 42: Selección de técnica de modelado Objetivo DM 1

4.1.2 Objetivo DM 2

Siguiendo el mismo modelo de tabla, la técnica y algoritmo para el Objetivo DM 2 cambia. En este caso primero se utilizará la técnica de *clustering* (agrupación) que generará varios conglomerados con tres de los atributos de los datos (Salvaje, Destructividad y Difusión) . Después, utilizando este agrupamiento obtenido se realizará la clasificación para saber si se han clasificado correctamente los datos restantes en cada grupo.

Nº	Objetivo DM	Técnica	Algoritmo
2	Obtener una clasificación para la clase "cluster" previamente generada	1. Clustering 2. Clasificación	1.1 EM 1.2 K-Means 2. J48(C4.5)

Tabla 43: Selección de técnica de modelado Objetivo DM 2

[7] Como se ha explicado anteriormente el algoritmo C4.5 se implementa en WEKA con el nombre J48.

4.1.3 Objetivo DM 3

Para obtener este objetivo se utiliza únicamente la técnica de clasificación, como en el objetivo de data mining 1, pero la complejidad está en obtener la clase id_ambito que se ha detallado más arriba , en el apartado 3.4 Integración de los datos.

Nº	Objetivo DM	Técnica	Algoritmo
3	Obtener una clasificación para la clase "id_ambito"	Clasificación	J48 (C4.5)

Tabla 44: Selección de técnica de modelado Objetivo DM 3

4.2 Generación de los Planes de Prueba

4.2.1 Plan de Prueba para el Objetivo 1 de Data Mining

Algoritmo J48	
Variable clase	Tipo de virus
Fuente de datos utilizada	T_VIRUS
Conjunto de entrenamiento	75 % de los datos de T_VIRUS
Conjunto de testeo	25 % de los datos de T_VIRUS
Modo de generar los conjuntos	Opción *Percentage Split* de WEKA
Medir la bondad	1. Porcentaje de instancias bien clasificadas 2. Precisión y cobertura del modelo

Tabla 45: Plan de prueba Objetivo DM 1

4.2.2 Plan de Prueba para el Objetivo 2 de Data Mining

Para la generación de esta prueba hay que distinguir entre la técnica de *clustering* que utiliza varios algoritmos y la técnica de clasificación que utiliza el algoritmo J48.

4.2.2.1 Plan de Prueba para el algoritmo de clustering EM utilizado en Objetivo de Data Mining 2

Algoritmo EM	
Fuente de datos utilizada	T_VIRUS
Conjunto de entrenamiento	Todos los datos
Conjunto de testeo	No hay
Medir la bondad	Comparar la variable log likelihood obtenida en diferentes ejecuciones del algoritmo con distintos valores de semilla.

Tabla 46: Plan de prueba algoritmo EM en Objetivo DM 2

4.2.2.2 Plan de Prueba para el algoritmo de clustering K-Means utilizado en el Objetivo de Data Mining2

Algoritmo K-Means	
Fuente de datos utilizada	T_VIRUS
Conjunto de entrenamiento	Todos los datos
Conjunto de testeo	No hay
Medir la bondad	Comparar la suma de errores cuadrado en los agrupamientos, obtenida en diferentes ejecuciones del algoritmo con distintos valores de semilla.

Tabla 47: Plan de prueba algoritmo K-Means en Objetivo DM 2

4.2.2.3 Plan de Prueba para el algoritmo de clasificación J48 utilizado en Objetivo de Data Mining 2

Algoritmo j48	
Variable clase	Tipo
Fuente de datos utilizada	T_VIRUS
Conjunto de entrenamiento	75 % de los registros de tabla T_VIRUS
Conjunto de testeo	25 % de los registros de la tabla T_VIRUS
Modo de generar los conjuntos	Opción *Percentage Split*de WEKA
Medir la bondad	1. Porcentaje de instancias bien clasificadas 2. Precisión y cobertura del modelo

Tabla 48: Plan de prueba algoritmo J48 en Objetivo DM 2

4.2.3 Plan de Prueba para el Objetivo 1.1 de Data Mining

Plan de Prueba para el Objetivo 3 de Data Mining	
Variable clase	Id_ambito
Fuente de datos utilizada	T_VIRUS
Conjunto de entrenamiento	75 % de los datos de T_VIRUS
Conjunto de testeo	25 % de los datos de T_VIRUS
Modo de generar los conjuntos	Opción *Percentage Split* de WEKA
Medir la bondad1	1. Porcentaje de instancias bien clasificadas 2. Precisión y cobertura del modelo

Tabla 49: Plan de prueba Objetivo DM 3

4.3 Construcción del Modelo

4.3.1 Descripción del modelo

Para cada construcción del modelo se describirá el(los) algoritmo(s) aplicado(s), descripción teórica sobre ellos y la configuración de parámetros utilizados en su ejecución.

4.3.1.1 Objetivo de Data Mining 1

Tal como se especifica en la *Tabla **42**: Selección de técnica de modelado Objetivo DM 1*la técnica de clasificación se aplica con el algoritmo clasificador J48.

4.3.1.1.1 Introducción teórica

Árboles de Clasificación

a) "Un clasificador es una partición del espacio de clasificación X en M subconjuntos disjuntos A1, A2,...,AM, siendo X[la unión de todos ellos y para todo x perteneciente a Am la clase predicha es Cm."

b) Tienen una interpretación muy fácil y sencilla dada su representación gráfica.

c) El procedimiento para generar el árbol consiste en seleccionar un atributo como nodo raíz(clase a predecir) y crear una rama con cada uno de los valores posibles de dicho atributo; con cada rama resultante se realiza el mismo proceso. En cada nodo se selecciona el atributo(selección de variable predictora) que mejor separe los ejemplos de acuerdo con la clase seleccionada.

d) Los árboles de decisión pueden ser fácilmente interpretados en forma de reglas "*si... entonces...*".

Algoritmo J48

a) El algoritmo J48 de WEKA es una implementación del algoritmo C4.5, uno de los algoritmos de minería de datos más utilizados en clasificación.

b) Resulta muy simple y potente como clasificador.

a) Se basa en la utilización del criterio ratio de ganancia (*gain ratio*) para la selección de la variable predictora en la subdivisión de la muestra. Es un proceso iterativo en el que cada submuestra se vuelve a dividir utilizando otra variable predictora.

c) El algoritmo incorpora una poda del árbol de clasificación una vez que éste ha sido construido.

Características específicas del algoritmo J48 de WEKA:

i. Los atributos que admite como variables predictoras pueden ser tanto numéricos como simbólicos (nominal).

ii. La clase o variable a predecir tiene que ser de tipo nominal.

iii. Admite atributos con valores *null* y *missing* tanto en el conjunto de entrenamiento como en la variable a predecir.

iv. Se permite ejemplos con peso.

v. Permite la realización del proceso de poda mediante la especificación del parámetro *reducedErrorPruning*[8]

[8] Es una opción específica del algoritmo en Weka, que indica si se realiza o no la poda del árbol en la ejecución del algoritmo.

4.3.1.1.2 Opciones de configuración para el algoritmo J48 en WEKA

Opción	Descripción
minNumObj (2)	Número mínimo de instancias por hoja
saveInstanceData (false)	Después de la creación del árbol de decisión se eliminan todas las instancias que se han clasificado en cada nodo, que han sido almacenadas previamente.
binarySplit (false)	No se divide cada nodo en dos ramas.
unpruned (False)	Se realiza la poda del árbol.
subtreeRaising(True)	Se permite realizar el podado con el proceso *Subtee raising* .
confidenceFactor (0.25)	Factor de confianza para el podado del árbol. Se utiliza el valor por defecto establecido por el algoritmo.
reducedErrorPruning(True)	El conjunto de ejemplos se divide en un subconjunto de entrenamiento y otro de test de los cuales el último servirá para estimar el error para la poda.
numFolds (2)	Define el número de subconjuntos en que hay que dividir el conjunto inicial para que el último de ellos se emplee como conjunto de test si se activa la opción *reducedErrorPruning*.
useLaplace (False)	No se emplea el suavizado de Laplace.

Tabla 50: Opciones de configuración para el algoritmo J48

4.3.1.2 Objetivo de Data Mining 2

Para lograr este objetivo se utilizan dos técnicas de data mining diferentes , clustering y clasificación.

4.3.1.2.1 Clustering

- Introducción Teórica

En esta técnica se han estudiado dos algoritmos(EM y K-Means) capaces de establecer automáticamente el número de conglomerados para poder ejecutar el algoritmo K-Means que obtendrá los centroides de estos conglomerados.

Cobweb

a) Es un algoritmo de clustering jerárquico.

b) Utiliza el aprendizaje incremental, realizando las agrupaciones instancia a instancia.

c) Durante la ejecución del algoritmo se forma un árbol de clasificación. Al principio, el árbol se forma con un único nodo raíz, consecutivamente las instancias se van añadiendo una a una y el árbol se va actualizando en cada paso. La actualización consiste en encontrar el mejor sitio donde incluir la nueva instancia, que puede necesitar de la reestructuración de todo el árbol.

d) Para la actualización del árbol el algoritmo utiliza la medida denominada *utilidad de categoría*, que mide la calidad general de una partición de instancias en un segmento. Para cada reestructuración se elige la que mayor utilidad de categoría proporcione.

Características específicas del algoritmo Cobweb en WEKA:

El algoritmo es muy sensible a dos parámetros:

a) **Acuity**: representa la medida de error de un nodo con una sola instancia, es decir, establece la varianza mínima de un atributo.

b) **Cut-off**: Este valor se utiliza para evitar el crecimiento desmesurado del número de segmentos. Indica el grado de mejoría que se debe producir en la utilidad de categoría para que la instancia sea tenida en cuenta de manera individual.

- Opciones de configuración para el algoritmo Cobweb en WEKA

Opción	Descripción
acuity (100)	Indica la mínima varianza permitida
cutoff(0.45)	Factor de poda.Indica la mejora en utilidad mínima por una subdivisión para que se permita llevar a cabo.

Tabla 51: Opciones de configuración para el algoritmo Cobweb

Aunque este algoritmo presenta muy buenas características de jerarquización y de gran utilidad para el problema planteado de conseguir un número de clusters adecuado, su fuerte dependencia de la variable cut-off ha impedido obtener, tras múltiples ejecuciones , un resultado óptimo e interpretable, por lo que finalmente no se ha utilizado.

EM (Expectation Maximization)

a) Es un algoritmo de agrupamiento por criterios estadísticos

b) Asigna a cada instancia una distribución de probabilidad de pertenencia a cada cluster.

c) El algoritmo puede decidir cuántos clusters crear o se le puede especificar a priori cuantos debe generar (en nuestro caso nos interesa que el algoritmo decida cuantos).

d) Utiliza el modelo Gaussiano finito de mezclas, asumiendo que todos los atributos son variables aleatorias independientes.

Opciones de configuración para el algoritmo EM en WEKA

Opción	Descripción
debug (false)	No se quiere mostrar la información sobre el proceso de clustering.
numClusters (-1)	El algoritmo determinará automáticamente el número de clusters.
maxIteration (100)	Número máximo de iteraciones del algoritmo. Se ha utilizado el valor por defecto que proporciona la herramienta.
seed(30)	Semilla a partir de la cual se generan los números aleatorios del algoritmo. Se han realizado varias pruebas con diferentes semillas. Los resultados se explicarán en el apartado 4.3.2 Objetivo de Data Mining 2
minStdDev (1.0E-6)	Factor de poda. Indica la mejora en utilidad mínima por una subdivisión para que se permita llevar a cabo. Se ha utilizado el valor por defecto que proporciona la herramienta.

Tabla 52: Opciones de configuración para el algoritmo EM

K-Means

a) Pertenece a la familia de algoritmos de particionado y recolocación(como el algoritmo EM)

b) Consiste en minimizar las distancias de los elementos de la partición y el centroide (media ponderada) de ésta.

c) Necesita la previa especificación del número de clusters.

Características específicas del algoritmo Cobweb en WEKA:

a) Admite atributos simbólicos y numéricos.

b) Para obtener los centroides iniciales se emplea un número aleatorio obtenido a partir del valor de la semilla especificado.

c) Los argumentos se normalizan.

4.3.1.3.4 Opciones de configuración para el algoritmo K-Means en WEKA

Opción	Descripción
numClusters (5)	Número de clusters (Este dato se obtiene a partir del algoritmo EM).
seed(10)	Semilla a partir de la cual se genera el número aleatorio para inicializar los centros de clusters. Se han realizado varias ejecuciones con valores de semillas diferentes. Los resultados se explicarán en el apartado 4.3.1.2 Objetivo de Data Mining 2

Tabla 53: Opciones de configuración para el algoritmo K-Means

4.3.1.2.2 Clasificación

Este modelo utiliza el algoritmo de clasificación J48 con los mismos parámetros tal como se explica en el apartado *Tabla **50**: Opciones de configuración para el algoritmo J48*.

4.3.1.3 Objetivo de Data Mining 3

Este modelo utiliza el algoritmo de clasificación J48 con los mismos parámetros tal como se explica en el apartado *Tabla **50**: Opciones de configuración para el algoritmo J48*.

4.3 Evaluación de los Modelos

4.3.1 Objetivo de Data Mining 1

De acuerdo con el plan de prueba para este objetivo (y para los restantes también), el conjunto de entrenamiento está formado por el 75% de los datos y el conjunto de prueba por los restante datos.

Activando la opción de *Percentage Split* de Weka sólo hace falta introducir el porcentaje de datos del conjunto de entrenamiento. Weka desordena los datos al azar utilizando la semilla previamente establecida en el campo *Random Seed for Xval* en Mas Opciones.

4.3.1.1 Resultado

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        1065               67.1078 %
Incorrectly Classified Instances       522               32.8922 %
Kappa statistic                          0.4026
Mean absolute error                      0.1149
Root mean squared error                  0.2452
Relative absolute error                 68.3375 %
Root relative squared error             84.4908 %
Total Number of Instances             1587

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0         0.003      0          0         0          0.727    Virus
                 0         0          0          0         0          0.814    VBS
                 0.672     0.055      0.512      0.672     0.581      0.914    Virus de macro
                 0.852     0.392      0.747      0.852     0.796      0.85     Gusano
                 0.479     0.15       0.538      0.479     0.507      0.82     Troyano
                 0         0          0          0         0          0.8      Hoax
                 0         0          0          0         0          0.861    otros
Weighted Avg.    0.671     0.27       0.614      0.671     0.639      0.841

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
   0   0  19  26  25   0   0 |   a = Virus
   0   0   4   2  12   0   0 |   b = VBS
   0   0  84  21  20   0   0 |   c = Virus de macro
   4   0  20 778 111   0   0 |   d = Gusano
   0   0  18 203 203   0   0 |   e = Troyano
   0   0   7   5   2   0   0 |   f = Hoax
   0   0  12   7   4   0   0 |   g = otros
```

Figura 11: Salida clasificador J48 para objetivo DM 1

4.3.1.2 Interpretación

Como se observa, el porcentaje de instancias bien clasificadas es el 67% de los datos de la muestra de prueba, formada con el 25 % de los datos iniciales (1587 instancias).

Por un lado, se puede observar que este modelo es muy deficiente para predecir las clases VBS, Hoax y Otros ya que todas sus instancias han sido mal clasificadas, pero por otro, en el caso de la clase Gusano, no lo es tanto. Los valores de precisión (*precision*) y de cobertura (*recall*) obtenidos para la clase Gusano son muy altos, mientras que para las clases VBS, Hoax son 0. El número total de instancias formadas por estas clases (VBS, Hoax y Otros) es de 55 (de la muestra del 25%), por lo que sólo representa el 3% del conjunto de prueba, pudiéndose así explicar el valor nulo obtenido en la variables de precisión y cobertura correspondientes a estas categorías. Para las clases predominantes y con mayor número de instancias respecto al conjunto en total de virus, sí se ha obtenido en su mayoría muy buen porcentaje de correctamente clasificados, generando así el resultado del 67,11% de virus bien clasificados para este modelo.

La matriz de confusión muestra la clasificación de las instancias. Las columnas indican la clasificación realizada por el clasificador y las filas recogen los datos reales. Por ejemplo, se observa que 111 elementos de la categoría Gusano han sido mal clasificados en la categoría de Troyanos. Las instancias bien clasificadas por el algoritmo forman la diagonal de la matriz. Aquí se observa claramente que todas las instancias de las clases VBS, Hoax y Otros han sido clasificadas en otras categorías.

Clasificadas								
Virus	VBS	Virus de macro	Gusano	Troyano	Hoax	Otros		Reales
0	0	19	26	25	0	0	Virus	
0	0	4	2	12	0	0	VBS	
0	0	84	21	20	0	0	Virus de macro	
4	0	20	778	111	0	0	Gusano	
0	0	18	203	203	0	0	Troyano	
0	0	7	5	2	0	0	Hoax	
0	0	12	7	4	0	0	Otros	

Tabla 54: Matriz de confusión objetivo DM 1

4.3.2 Objetivo de Data Mining 2

Para cumplir con este objetivo de mining se utilizan dos técnicas diferentes, clustering e clasificación.

4.3.2.1 Clustering

Primero, se agruparán los datos en un número concreto de clusters, utilizando para ello el algoritmo EM, que es capaz de establecer automáticamente el número de clusters. Después, se utilizará el algoritmo K-MEANS, que obtendrán los centroides[9] de cada conglomerado , con el fin de conocer características comunes de cada grupo. El clustering se ejecutará sobre tres de los ocho atributos de la tabla T_VIRUS (Salvaje, Destructividad y Difusión).

Algoritmo EM

El valor de la semilla que se utiliza en la ejecución del algoritmo EM tiene suma importancia puesto que, para las diferentes ejecuciones que se han realizado, el número de clusters que devuelve el algoritmo no es el mismo.

Se ha ejecutado varias veces el algoritmo EM con diferentes semillas aleatorias. Para medir la bondad de dicho algoritmo se utiliza la variable *log likelihood*, que se basa en valores de la función de densidad y se puede interpretar en términos comparativos con otras ejecuciones como "a mayor valor obtenido mejor es el resultado del algoritmo". Se ha obtenido el mayor valor de esta variable con la semilla aleatoria 30 y se han obtenido 5 diferentes clusters, que se utilizarán para ejecutar el algoritmo K-Means.

```
Time taken to build model (full training data) : 70.52 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       152 (  2%)
1       824 ( 13%)
2      2052 ( 32%)
3      2225 ( 35%)
4      1094 ( 17%)


Log likelihood: -2.633
```

Figura 12: Salida algoritmo EM para objetivo DM 2

[9] Punto que define el centro geométrico

Algoritmo K-Means

Para obtener los 5-centroides iniciales, donde 5 es número de clusters establecido por algoritmo EM, K-Means emplea el valor de la semilla fijado previamente, por lo que es importante establecer el mejor valor posible para esta variable. Para ello, se han realizado varias ejecuciones del algoritmo comparando la suma de errores cuadrados en las agrupaciones creadas por el algoritmo.

El mejor resultado (menor número de suma de errores cuadrados) es el siguiente:

Clusterer output

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 3839.0
Missing values globally replaced with mean/mode

Cluster centroids:
                            Cluster#
Attribute          Full Data          0          1          2          3          4
                      (6347)     (3331)     (1211)      (566)      (145)     (1094)
=============================================================================================
SALVAJE                   MB         MB         MB         MB         MB Desconocido
DESTRUCTIVIDAD             M          M          A          M         MB Desconocido
DIFUSION                   B          B          M          A         MB Desconocido



Time taken to build model (full training data) : 0.03 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       3331 ( 52%)
1       1211 ( 19%)
2        566 (  9%)
3        145 (  2%)
4       1094 ( 17%)
```

Figura 13: Salida algoritmo K-Means para objetivo DM 2

En la figara *Figura **13*** se reflejan los centroides para cada atributo de cada uno de los 5 clusters (de 0 a 4).

Si damos prioridad al atributo DESTRUCTIVIDAD, el Cluster# 1 definiría el conglomerado de virus con más altos valores (Destructividad alta (A) ,Difusión media (M), Salvaje muy bajo (MB)). Una característica destacable es, que el atributo salvaje sólo toma valores *desconocido* y MB (muy bajo) en todos los conglomerados.

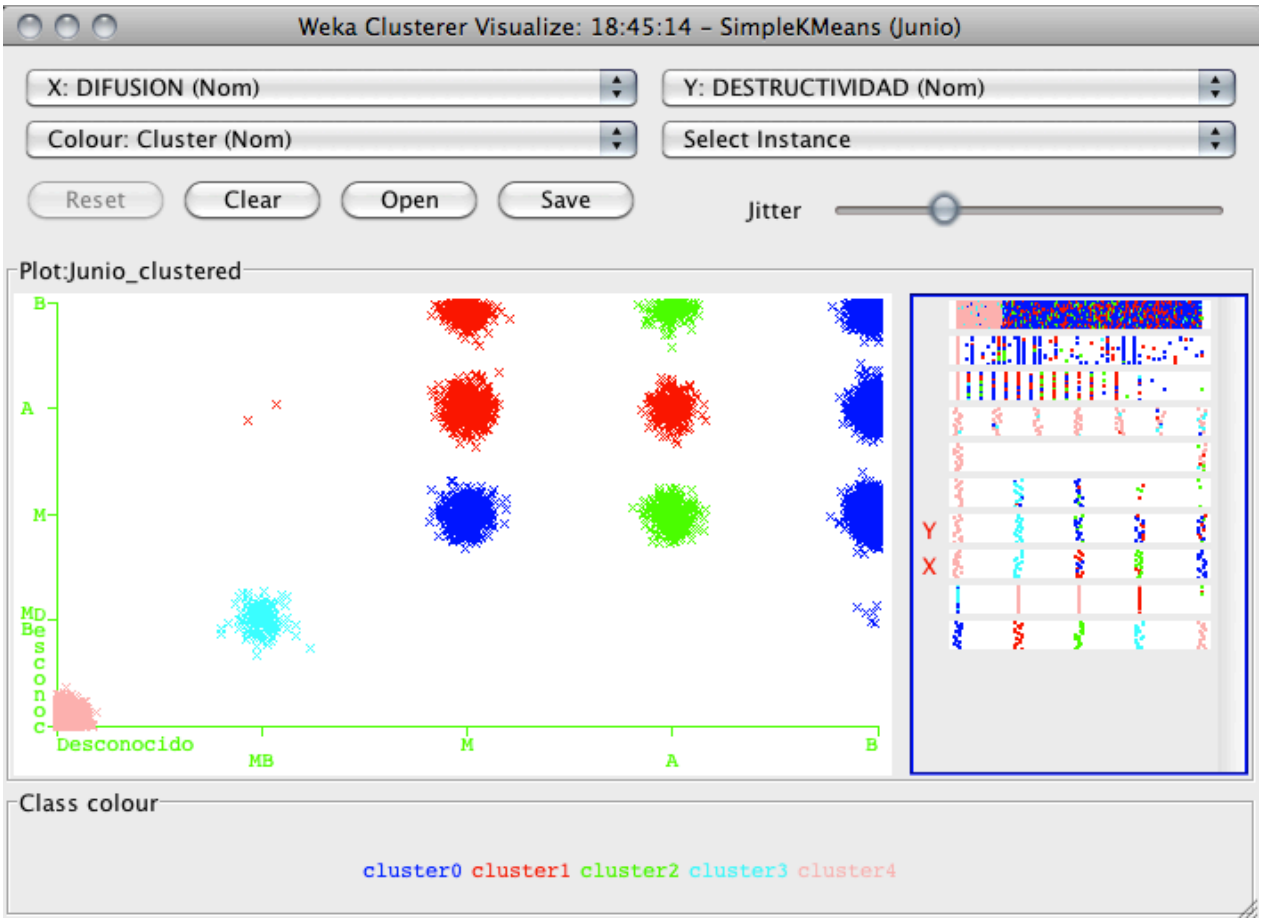

Figura 14: Visualización cluster K-Means para objetivo DM 2

Sin embargo, representando gráficamente los clusters (en WEKA sólo se pueden representar gráficas 2D , eje X, eje Y y una clase con color), observamos que el mismo grupo 1 (clase cluster1 con color rojo), aunque represente de media ponderada la categoría de Destructividad con valor alto y Difusión con valor medio, también agrupa los virus con Destructividad media y Difusión alta y los virus con Destructividad media y Difusión media. Dicho esto, podemos decir que el conglomerado #1 representa los virus más dañinos o peligrosos.

Interpretando los demás resultados de la gráfica junto con los centroides de cada grupo, se puede afirmar que, como se ha mencionado anteriormente, los virus más dañinos son los del cluster1 (rojo), seguidos del cluster 2 (verde), el cluster0 (azul) , el cluster 3 (turquesa) y, por último, el cluster 4 (rosa).

4.3.2.2 Clasificación

Conociendo las características de cada uno de los cinco grupos establecidos por el algoritmo K-Means, se ejecuta el algoritmo de clasificación J48 sobre los restantes 5 atributos (NOM_PROPAGACION, NOM_PLATAFORMA, TIPO, DESTACADO, PELIGRO), para conseguir un modelo de clasificación para los virus de cada cluster.

Se han utilizado los mismos parámetros de ejecución que se utilizaron para el Objetivo DM 1 y se obtiene el resultado que se muestra a continuación.

4.3.2.1 Resultado

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        1166               73.472  %
Incorrectly Classified Instances       421               26.528  %
Kappa statistic                          0.5846
Mean absolute error                      0.1547
Root mean squared error                  0.2791
Relative absolute error                 59.4618 %
Root relative squared error             77.318  %
Total Number of Instances             1587

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.834     0.231      0.8       0.834     0.817      0.856    cluster1
                 0.535     0.097      0.566     0.535     0.55       0.821    cluster2
                 0.413     0.013      0.765     0.413     0.537      0.866    cluster3
                 0.308     0.005      0.6       0.308     0.407      0.757    cluster4
                 0.897     0.072      0.71      0.897     0.793      0.959    cluster5
Weighted Avg.    0.735     0.153      0.732     0.735     0.725      0.865

=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 695  81   5   8  44 |   a = cluster1
 111 162  14   0  16 |   b = cluster2
  20  37  62   0  31 |   c = cluster3
  19   3   0  12   5 |   d = cluster4
  24   3   0   0 235 |   e = cluster5
```

Figura 15:Salida clasificador J48 para el objetivo DM 2

4.3.2.2 Interpretación

El porcentaje de instancias bien clasificadas del conjunto de prueba, que contiene el 25% de los datos, es de 73%, consiguiendo así llegar al objetivo de 70 % de instancias bien clasificadas.

En este caso no se obtiene ningún valor 0 en las medidas de precisión y cobertura (recall) de ninguna de las categorías de virus. La media de porcentaje de todas ellas es de 0,732 en

precisión y 0,735 en cobertura, mejorando así considerablemente la predicción obtenida respecto al Objetivo DM 1.

La *Figura **15*** incluye la matriz de confusión. Se observa, además, que la variable cluster tiene 5 clases , empezando por el cluster 1 hasta el cluster 5. Esto es debido a que el algoritmo K-Means empieza la enumeración de los cluster con 0 y el clasificado J48 con el valor 1, por lo que el grupo de virus más dañino, tal como se explicó más arriba, era el cluster 1, y en los resultados que se muestran en la *Figura **15*** y la *Figura **16*** es el cluster 2.

Comprobada así la bondad del modelo de clasificación obtenido, podemos tomar como válidas algunas reglas obtenidas por el clasificador, como por ejemplo:

1) Los virus con NOM_PLATAFORMA >0 , NOM_PROPAGACION > 1 y PELIGRO > 2 pertenecen al cluster 2.

2) Los virus con NOM_PLATAFORMA >0 , NOM_PROPAGACION > 1 , PELIGRO <= 0, y NOM_PLATAFORMA <= 8 pertenecen al cluster 2.

3) Los virus con NOM_PLATAFORMA >0 , NOM_PROPAGACION <= 1 , PELIGRO > 1 y <=2 y TIPO= VIRUS pertenecen al cluster 1.

4) Los virus con NOM_PLATAFORMA >0 , NOM_PROPAGACION <= 1 y PELIGRO > 2 pertenecen al cluster 2.

Como ya sabemos que el grupo de virus más dañino es el cluster 2 en la clasificación (cluster 1 en lo resultados del algoritmo K-Means) e interpretando, por ejemplo,la última regla citada más arriba, podemos deducir que lo virus que forman el grupo más dañino su NOM_PLATAFORMA es mayor que 0, NOM_PROPAGACIÓN es menor o igual a 1 y PELIGRO es mayor que 2.

```
NOM_PLATAFORMA > 0
|   NOM_PROPAGACION <= 1
|   |   PELIGRO <= 1: cluster1 (1439.0/162.0)
|   |   PELIGRO > 1
|   |   |   PELIGRO <= 2
|   |   |   |   TIPO = Virus: cluster1 (3.0)
|   |   |   |   TIPO = VBS: cluster1 (0.0)
|   |   |   |   TIPO = Virus de macro: cluster1 (4.0)
|   |   |   |   TIPO = Gusano
|   |   |   |   |   NOM_PROPAGACION <= 0
|   |   |   |   |   |   NOM_PLATAFORMA <= 21: cluster3 (82.0/46.0)
|   |   |   |   |   |   NOM_PLATAFORMA > 21: cluster1 (4.0/1.0)
|   |   |   |   |   NOM_PROPAGACION > 0: cluster2 (2.0/1.0)
|   |   |   |   TIPO = Troyano: cluster1 (76.0/9.0)
|   |   |   |   TIPO = Hoax: cluster1 (0.0)
|   |   |   |   TIPO = otros: cluster1 (1.0)
|   |   |   PELIGRO > 2: cluster2 (12.0/1.0)
|   NOM_PROPAGACION > 1
|   |   PELIGRO <= 1
|   |   |   PELIGRO <= 0
|   |   |   |   NOM_PLATAFORMA <= 8: cluster2 (20.0/8.0)
|   |   |   |   NOM_PLATAFORMA > 8
|   |   |   |   |   NOM_PROPAGACION <= 7: cluster2 (524.0/281.0)
|   |   |   |   |   NOM_PROPAGACION > 7
|   |   |   |   |   |   NOM_PROPAGACION <= 8: cluster2 (29.0/3.0)
|   |   |   |   |   |   NOM_PROPAGACION > 8
|   |   |   |   |   |   |   NOM_PLATAFORMA <= 31
|   |   |   |   |   |   |   |   NOM_PROPAGACION <= 9: cluster1 (46.0/25.0)
|   |   |   |   |   |   |   |   NOM_PROPAGACION > 9: cluster2 (100.0/28.0)
|   |   |   |   |   |   |   NOM_PLATAFORMA > 31: cluster1 (5.0/1.0)
|   |   |   PELIGRO > 0: cluster1 (135.0/48.0)
|   |   PELIGRO > 1
|   |   |   PELIGRO <= 2: cluster3 (173.0/45.0)
|   |   |   PELIGRO > 2: cluster2 (48.0/2.0)

Number of Leaves  :     32

Size of the tree :      53


Time taken to build model: 0.11 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances        1166               73.472  %
Incorrectly Classified Instances       421               26.528  %
Kappa statistic                          0.5846
Mean absolute error                      0.1547
Root mean squared error                  0.2791
Relative absolute error                 59.4618 %
Root relative squared error             77.318  %
Total Number of Instances             1587
```

Figura 16: Reglas del clasificador J48 para el objetivo DM 2

4.3.3 Objetivo de Data Mining 3

Para lograr este objetivo se ha ejecutado el algoritmo J48 con la misma configuración de parámetros que la utilizada para el objetivo DM 1, utilizando como atributo de clase la variable id_ambito.

El resultado que se ha obtenido, como se observa en figura, es muy malo; sólo el 27% de las instancias han sido bien clasificadas.

Classifier output

```
=== Summary ===

Correctly Classified Instances         721               27.7949 %
Incorrectly Classified Instances      1873               72.2051 %
Kappa statistic                          0.048
Mean absolute error                      0.224
Root mean squared error                  0.3352
Relative absolute error                 97.7571 %
Root relative squared error             99.0511 %
Total Number of Instances             2594

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.463     0.276       0.402     0.463      0.43       0.61     1
                 0         0.002       0         0          0          0.515    2
                 0.662     0.671       0.218     0.662      0.328      0.491    3
                 0         0           0         0          0          0.625    4
                 0         0           0         0          0          0.538    5
                 0         0           0         0          0          0.557    6
                 0         0           0         0          0          0.596    7
Weighted Avg.    0.278     0.227       0.163     0.278      0.195      0.552

=== Confusion Matrix ===

   a   b   c   d   e   f   g   <-- classified as
 343   5 393   0   0   0   0 |   a = 1
 107   0 277   0   0   0   0 |   b = 2
 193   0 378   0   0   0   0 |   c = 3
  22   0  97   0   0   0   0 |   d = 4
 123   0 345   0   0   0   0 |   e = 5
  61   0 217   0   0   0   0 |   f = 6
   5   0  28   0   0   0   0 |   g = 7
```

Figura 17: Salida clasificador J48 para el objetivo DM 2

A modo de prueba se ha ejecutado el algoritmo utilizando la misma variable a predecir que se utilizó para cubrir el objetivo DM 1 (tipo de virus) y se ha obtenido un resultado inesperado.

4.3.3.1 Resultado

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         515               79.4753 %
Incorrectly Classified Instances       133               20.5247 %
Kappa statistic                          0.5807
Mean absolute error                      0.0846
Root mean squared error                  0.2127
Relative absolute error                 50.6455 %
Root relative squared error             73.4563 %
Total Number of Instances              648

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.318     0.008       0.583     0.318     0.412      0.945    Virus
                 0         0           0         0         0          0.741    VBS
                 0.866     0.113       0.526     0.866     0.654      0.926    Virus de macro
                 0.417     0.037       0.476     0.417     0.444      0.909    Troyano
                 0.935     0.172       0.922     0.935     0.929      0.954    Gusano
                 0.025     0.012       0.125     0.025     0.042      0.903    otros
Weighted Avg.    0.795     0.136       0.763     0.795     0.77       0.94

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
   7   0   8   2   4   1 |   a = Virus
   0   0   4   2   3   2 |   b = VBS
   1   0  71   8   2   0 |   c = Virus de macro
   0   0   3  20  23   2 |   d = Troyano
   4   0  13  10 416   2 |   e = Gusano
   0   0  36   0   3   1 |   f = otros
```

Figura 18: Salida clasificador J48 para objetivo DM 3 con variable clase "tipo de virus"

4.3.3.2 Interpretación

El porcentaje de instancias bien clasificadas es muy alto, casi el 80%, mostrando una media de precisión (*precision*) y cobertura (*recall)* que se acerca también al 80%.

Se observa también que para la clase Gusano dichos valores de precisión y cobertura superan el 90%, siendo la clase que más cobertura ha tenido para este modelo, pero para las categorías de VBS y Otros las variables precisión y cobertura muestran valores muy bajos, incluso 0. Estos valores, tal como se ha explicado en los modelos anteriores, se pueden deber a que el número de virus de pertenecientes a estas dos categorías representa un porcentaje muy pequeño del conjunto de prueba.

A partir de la figura que viene a continuación, se pueden interpretar algunas de las reglas obtenidas por el algoritmo clasificador que se pueden utilizar para predecir qué virus actúan en un ámbito (sector de institución) concreto.

- Si nom_plataforma = 9 y Difusión = Baja y Destructividad = Media y Id_ambito = 1 entonces el virus es Gusano
- Si nom_plataforma = 9 y Difusión = Baja y Destructividad = Media y Id_ambito = 24 entonces el virus es Troyano
- Si nom_plataforma = 9 y Difusión = Baja y Destructividad = Baja entonces el virus es Gusano

```
NOM_PLATAFORMA = 9
|   DIFUSION = Desconocido: Gusano (0.0)
|   DIFUSION = MB: Gusano (6.0)
|   DIFUSION = A: Gusano (643.0/1.0)
|   DIFUSION = M
|   |   NOM_PROPAGACION = 0
|   |   |   PELIGRO = 0: Gusano (18.0/3.0)
|   |   |   PELIGRO = 1: Gusano (11.0/6.0)
|   |   |   PELIGRO = 2: Troyano (7.0)
|   |   |   PELIGRO = 3: Gusano (0.0)
|   |   |   PELIGRO = 4: Gusano (0.0)
|   |   NOM_PROPAGACION = 1: Gusano (3.0)
|   |   NOM_PROPAGACION = 2: Gusano (184.0)
|   |   NOM_PROPAGACION = 3: Gusano (0.0)
|   |   NOM_PROPAGACION = 4: Gusano (17.0)
|   |   NOM_PROPAGACION = 5: Gusano (2.0)
|   |   NOM_PROPAGACION = 6: Gusano (1.0)
|   |   NOM_PROPAGACION = 7: Gusano (26.0)
|   |   NOM_PROPAGACION = 8: Gusano (0.0)
|   |   NOM_PROPAGACION = 9: Gusano (1.0)
|   |   NOM_PROPAGACION = 11: Gusano (1.0)
|   |   NOM_PROPAGACION = 13: Gusano (0.0)
|   DIFUSION = B
|   |   DESTRUCTIVIDAD = Desconocido: Gusano (0.0)
|   |   DESTRUCTIVIDAD = MB: Gusano (0.0)
|   |   DESTRUCTIVIDAD = A: Troyano (30.0/18.0)
|   |   DESTRUCTIVIDAD = M
|   |   |   id_ambito = 1: Gusano (13.0/6.0)
|   |   |   id_ambito = 2: Troyano (10.0/3.0)
|   |   |   id_ambito = 3: Gusano (17.0/6.0)
|   |   |   id_ambito = 4: Troyano (2.0)
|   |   |   id_ambito = 5: Troyano (17.0/9.0)
|   |   |   id_ambito = 6: Gusano (5.0/2.0)
|   |   |   id_ambito = 7: Troyano (1.0)
|   |   DESTRUCTIVIDAD = B: Gusano (35.0/15.0)
NOM_PLATAFORMA = 12: Virus de macro (5.0/2.0)
NOM_PLATAFORMA = 13: Gusano (0.0)
NOM_PLATAFORMA = 15
|   PELIGRO = 0: Gusano (13.0)
|   PELIGRO = 1
|   |   DESTRUCTIVIDAD = Desconocido: Troyano (0.0)
|   |   DESTRUCTIVIDAD = MB: Troyano (0.0)
|   |   DESTRUCTIVIDAD = A: Troyano (0.0)
|   |   DESTRUCTIVIDAD = M: Troyano (2.0)
|   |   DESTRUCTIVIDAD = B: Gusano (2.0)
```

Figura 19: Reglas del clasificador J48 para objetivo DM 3 con variable clase "tipo de virus"

5. EVALUACIÓN

5.1 Evaluación de los resultados

La evaluación exhaustiva de los resultados y la interpretación de cada uno de los modelos se ha llevado a cabo en su correspondiente sección de evaluación de los modelo, 4.3 Evaluación de los Modelos.

A continuación, para comprender mejor la satisfacción de cada uno de los objetivos de negocio identificados, se muestra primero la tabla de valoración de cada uno de los objetivos de data mining y su criterio de éxito, y después los resultados logrados para cada objetivo de negocio.

Nº	Objetivo DM	Medida de Éxito	Logrado
1	Obtener una clasificación para la clase "tipo" de virus	Obtener un modelo de clasificación con un conjunto de entrenamiento formado por el 75 % de los datos de la fuente de datos T_VIRUS y un conjunto de testeo con el 25 % restante y lograr un porcentaje de éxito mayor del 70% y resultados óptimos en las medidas de precisión y cobertura.	No. No se ha conseguido un buen resultado en la clasificación de todas las categorías de la clase "tipo". Por lo tanto, este modelo se toma como deficiente.
2	Obtener una clasificación para la clase "cluster" previamente generada	Obtener un modelo de clasificación con un conjunto de entrenamiento formado por el 75 % de los datos de la fuente de datos T_VIRUS y un conjunto de testeo con el 25 % restante y lograr un porcentaje de éxito mayor del 70% y resultados óptimos en las medidas de precisión y cobertura.	Sí. Se ha obtenido un porcentaje mayor de 70% de instancias bien clasificados, consiguiendo también buenos valores para las variables de precisión y cobertura que comprueban la bondad del modelo.
3	Obtener una clasificación para la clase "id_ámbito" utilizando la fuente de datos T_VIRUS2	Obtener un modelo de clasificación con un conjunto de entrenamiento formado por el 75 % de los datos de la fuente de datos T_VIRUS y un conjunto de testeo con el 25 % restante y lograr un porcentaje de éxito mayor del 70% y resultados óptimos en las medidas de precisión y cobertura.	Por un lado, no se ha conseguido un buen resultado en la predicción de la variable id_ámbito (sólo 30% bien clasificados), pero por otro, sí se ha conseguido un alto porcentaje en instancias de tipo virus bien clasificadas(79%) utilizando la fuente de datos T_VIRUS2 que contiene información sobre los virus que atacan a cada sector de institución.

Tabla 55: Evaluación final de los objetivos DM

Objetivos de Negocio

	Nº 1 Conocer características de todos los virus interceptados con el fin de predecirlos.	**Nº 2** Conocer información relevante sobre los virus que atacan a cada sector de institución para poder tomar medidas de prevención frente a ellos.
Medida de éxito	Obtener información útil para poder prevenir los ataques.	Obtener información relevante de los virus que atacan a cada ámbito de institución.
Objetivo DM correspondiente	Objetivo DM 1 y DM 2	Objetivo DM 3
¿Logrado?	No, sólo se ha podido lograr el objetivo de DM 2, por lo que finalmente no podemos satisfacer la medida de éxito para este objetivo de negocio. Además, tampoco podemos asegurar una predicción para todos los tipos de virus.	S , se ha logrado este objetivo de negocio. Se ha tomado como decisión final tomar como referencia los resultados obtenidos del clasificador con la variable clase “tipo de virus ”, puesto que se han obtenido reglas válidas que puedan ser utilizadas para predecir qué virus puede atacar a un ámbito de institución concreto *Figura **19**: Reglas del clasificador J48para objetivo DM 3 con variable clase “tipo de virus”*

Tabla 56: Evaluación final de los objetivos de negocio

5.2 Revisión del proceso

La revisión del proceso sigue la línea de desarrollo tal como se ha implementado en el proyecto (basándose en la metodología CRISP-DM), identificando las seis fases del análisis de datos.

5.2.1 Comprensión del problema

Actividades	Descripción
Tareas realizadas	Determinar los objetivos del negocio. Establecer los criterios de éxito correspondiente a cada objetivo del negocio. Realizar el inventario de recursos. Establecer de requisitos, suposiciones y restricciones. Identificar los riesgos y plan de contingencia. Listar la terminología de negocio y de Data Mining. Determinar los objetivos de Data Mining correspondientes a los objetivos de negocio. Establecer los criterios de éxito correspondiente a cada objetivo DM. Crear el plan de proyecto correspondiente.
¿Qué fue necesario?	Todas las tareas realizadas fueron necesarias.
¿Qué fue realizado óptimamente?	Determinar los objetivos del negocio. Identificar los riesgos y plan de contingencia .
¿Qué se puede mejorar en líneas futuras?	Decidir implementar el software. Para ello se incorporarían más objetivos de negocio y, por consiguiente, el desarrollo del análisis sería de mayor volumen y más complejo. Contar con un especialista en la fuente de datos, para que pueda proporcionar información más precisa.
Fallos encontrados	
Pasos desviados	
Acciones alternativas	

Tabla 57: Revisión de la fase: Comprensión del problema

5.2.2 Comprensión de los datos

Actividades	Descripción
Tareas realizadas	Recoger y detallar todas las fuentes de datos. Describir las fuentes de datos utilizadas y sus valores en atributos. Realizar el informe de exploración de datos para las fuentes de datos que se van a utilizar. Realizar el informe de verificación de la Calidad de los Datos para las fuentes seleccionadas.
¿Qué fue necesario?	Todas las tareas realizadas en esta fase son necesarias.
¿Qué fue realizado óptimamente?	Describir las fuentes de datos utilizadas y sus valores en los atributos. El informe de exploración de datos para las fuentes de datos - Informe de calidad de datos. El informe de verificación de la Calidad de los Datos para las fuentes seleccionadas.
¿Qué se puede mejorar en líneas futuras?	Contar con un especialista de la fuente de datos para mejorar la descripción de todas las fuentes de datos y sus valores, la exploración de datos, el informe de verificación de calidad de datos. Disponer de una fuente de datos más limpia y balanceada.
Fallos encontrados	Hay demasiados valores null en atributos. Hay atributo sólo con valores null. Hay valores missing en atributos. Hay valores que no se corresponden con los valores permitidos del atributo.
Pasos desviados	
Acciones alternativas posibles	

Tabla 58: Revisión de la fase: Comprensión de los datos

5.2.3 Preparación de los datos

Actividades	Descripción
Tareas realizadas	Seleccionar de las fuentes de datos que se utilizarán Seleccionar los atributos necesarios de cada fuente de datos. Realizar la limpieza de los datos seleccionados. Construir los datos necesarios. Realizar diferentes combinaciones entre tablas para integrar los datos. Formatear los datos.
¿Qué fue necesario?	Todas las tareas realizadas en esta fase tienen suma importancia.
¿Qué fue realizado óptimamente?	Seleccionar los atributos necesarios de cada fuente de datos para la combinación posterior de los datos. Limpieza de los datos utilizados. Combinar apropiadamente varias fuentes de datos para integrar los datos.
¿Qué se puede mejorar en líneas futuras?	Mejorar la limpieza de los datos basándose en más medidas de precisión. Disponer de un especialista en la fuente de datos para realizar las tareas de limpieza, transformación y combinación de datos.
Fallos encontrados	
Pasos desviados	
Acciones alternativas posibles	

Tabla 59: Revisión de la fase: Preparación de los datos

5.2.4 Modelado

Actividades	Descripción
Tareas realizadas	Seleccionar las técnicas de modelado. Generar los modelos para cada uno de los objetivos DM. Construir los modelos de datos establecidos. Evaluar cada uno de los modelos establecidos.
¿Qué fue necesario?	Todas las tareas realizadas fueron necesarias.
¿Qué fue realizado óptimamente?	Seleccionar las técnicas de modelado para cada uno de los objetivos de DT. Construir cada uno de los modelos, detallando aspectos técnicos y de configuración para los algoritmos elegidos. Evaluar exhaustivamente las salidas de cada modelo.
¿Qué se puede mejorar en líneas futuras?	Mejorar los conjuntos de entrenamiento y de testeo con datos más balanceados.
Fallos encontrados	Modelo deficiente para el objetivo DM 1 y para la elección de la clase a predecir en el objetivo DM 3.
Pasos desviados	Utilizar la variable *tipo* de virus como clase a predecir en el modelo correspondiente al objetivo DM 3.
Acciones alternativas posibles	

Tabla 60: Revisión de la fase: Modelado

5.2.5 Evaluación de los modelos

Actividades	**Descripción**
Tareas realizadas	Evaluación general de los resultados obtenidos de cada modelo y su correspondiente criterio de éxito. Revisar el proceso. Líneas futuras del proyecto.
¿Qué fue necesario?	Todas las tareas realizadas fueron necesarias.
¿Qué fue realizado óptimamente?	Evaluar los resultados obtenidos de cada modelo y su correspondiente criterio de éxito. Revisar todo el proceso. Detallar posibles líneas futuras del proyecto.
¿Qué se puede mejorar en líneas futuras?	
Fallos encontrados	
Pasos desviados	
Acciones alternativas posibles	

Tabla 61: Revisión de la fase: Evaluación de los modelos

5.3 Líneas futuras

Al ser limitado el tiempo dedicado a este proyecto (las 324 horas correspondientes a los 12 ECTS de la asignatura) han quedado fuera del alcance del proyecto algunas ideas que hubieran podido resultar interesantes, algunas de las cuales se recogen en la tabla siguiente como posibles líneas futuras del proyecto.

Nº	Tarea	Propósito
1	Desarrollar un sistema software de predicción de virus.	La meta del análisis de datos es extraer conocimiento nuevo e interesante para los objetivos del negocio. En el caso del presente proyecto, el conocimiento descubierto podría ser utilizado para el desarrollo de un sistema software de predicción de ataques de virus.
2	Realizar una nueva iteración completa del proceso KDD guiada por un experto en el dominio de los datos.	Una búsqueda a ciegas no garantiza que se alcance con éxito el objetivo del KDD. Es importante conocer a fondo el dominio del problema y entender a fondo los datos para poder guiar el proceso KDD y así tomar las decisiones acertadas en cada uno de los pasos de dicho proceso. Contando con la colaboración de un experto en el dominio del problema se podría realizar una nueva iteración del proceso KDD para corregir errores cometidos y precisar aún más todas las acciones llevadas a cabo, con el fin de obtener resultados más óptimos.
3	Utilizar otros algoritmos clasificadores de data mining.	Sería interesante probar otros algoritmo de data mining, como por ejemplo, las redes de neuronas. Las redes de neuronas son muy buenos clasificadores una vez que están entrenadas, pero todo entrenamiento para ajustar los pesos de cada neurona, y así como las pruebas que hay que realizar para elegir la mejor arquitectura de la red, llevan su tiempo de dedicación, y como hemos señalado, el tiempo de desarrollo de este trabajo es limitado.

Tabla 62: Líneas futuras del proyecto

6. DESPLIEGUE

Esta fase no se desarrollará por quedar fuera del alcance de este proyecto.

VI. RESULTADOS

Se resumen e identifican los resultados obtenidos con la realización del proyecto, siguiendo el orden establecido para cada uno de los 10 objetivos.

Nº obj.	Descripción del objetivo	¿Logrado?
1	Definir el tipo de conocimiento que se espera encontrar	Sí
2	Definir el algoritmo de data mining a utilizar	Sí
3	Aprender a utilizar y conocer el proceso software "*CRISP-DM v1.0*"	Sí
4	Aprender a utilizar y conocer la herramienta "*WEKA*"	Sí
5	Realizar la selección de los datos	Sí
6	Realizar la preparación y limpieza de los datos	Sí
7	Realizar la transformación de los datos	Sí
8	Definir (o establecer) los parámetros con que se ejecutará el algoritmo de data mining a aplicar	Sí
9	Modelar los datos	Sí
10	Evaluar los resultados obtenidos	Sí

Tabla 63: Objetivos logrados

1. Definir el tipo de conocimiento que se espera encontrar

A pesar de que el proyecto finalmente no fue confirmado por parte del cliente, se establecieron claramente dos tipos de objetivos de negocio (*Tabla 2: Objetivos del negocio*) intentando que éstos imitasen y se ajustasen de la mejor forma posible a la idea inicial de la implementación del software de predicción de virus.

2. Definir el algoritmo de data mining a utilizar

Para cada uno de los objetivos definidos de data mining, se eligió el algoritmo que mejor satisficiera sus criterios de éxito. Los algoritmos que se han utilizado son J48 (es un algoritmo clasificador), EM y K-Means (ambos algoritmos son de agrupamiento).

3. Aprender a utilizar y conocer el proceso software "CRISP-DM"

Para seguir un desarrollo estructurado de Data Mining se ha tomado como referencia la metodología "*CRISP-DM*", plasmando todas las actividades y salidas necesarias para la comprensión de cada una de sus 6 fases definidas.

4. Aprender a utilizar y conocer la herramienta "WEKA"

Se ha aprendido a utilizar y conocer en profundidad esta herramienta de minería de datos. Weka es un software de libre distribución y se puede encontrar mucha información y tutoriales que detallan su utilización. Debido a los algoritmos empleados en este trabajo, se ha profundizado con los paneles de pre-procesamiento, clustering, visualización y clasificación que proporciona la herramienta en la interfaz Explorer.

5. Realizar la selección de los datos

Se ha realizado de manera óptima la selección de los datos, realizando un previo estudio de cada una de las fuentes de datos iniciales, excluyendo la información no relevante para el análisis de datos.

6. Realizar la preparación y limpieza de los datos

Para la preparación y limpieza de los datos se ha elaborado el correspondiente informe de exclusión/inclusión tanto para fuentes de datos como para atributos de las tablas seleccionadas, con el fin de utilizar sólo información útil en el proceso. También se han eliminado algunos registros con variables que están fuera de los rangos permitidos para ese atributo.

7. Realizar la transformación de los datos

Se han realizado dos acciones de transformación: transformar valores de algunos atributos para no perder esos registros, y combinar información de varias fuentes de datos, para crear nuevos registros.

8. Definir (o establecer) los parámetros del algoritmo de data mining que se van a aplicar

Se han definido, establecido y explicado las diferentes opciones de parámetros permitidas en las ejecuciones de los algoritmos utilizados.

9. Modelar los datos

Para la fase de modelado de los datos se han seleccionado las técnicas de modelado, se han generado los correspondientes planes de prueba y se ha construido el modelo completo definido para cada objetivo DM necesario en este trabajo.

10. Evaluar los resultados obtenidos

Se han evaluado todos los resultados obtenidos tras las ejecuciones de los algoritmos, mostrando claramente su valoración respecto a los objetivos de data mining y respecto a los objetivos de negocio concretos para este trabajo.

VII. CONCLUSIONES

Tras el desarrollo de este proyecto se ha llegado a las siguientes conclusiones:

I. La metodología utilizada, CRISP-DM, es un notable referente para proyectos de minería de datos que proporciona una guía muy completa para llevar a cabo el trabajo en toda su profundidad. Siendo esta metodología muy detallada en todas sus fases, permite que un proyecto de esta modalidad pueda ser fácilmente comprendido por una persona sea o no especialista en data mining, en particular, y KDD en general.

II. El tratamiento realizado sobre los datos (análisis, selección, transformación y limpieza) es muy relevante, y muy determinante, en el proceso de KDD. Es importante disponer de información descriptiva y esclarecedora sobre toda la fuente de datos, que permita tomar las decisiones adecuadas en dicho tratamiento de los datos. Sin este entendimiento de los datos y del dominio del problema es fácil dar un mal paso en cualquiera de las tareas relacionadas con el tratamiento de los datos, lo cual puede provocar que se obtengan resultados erróneos.

III. Las técnicas empleadas en este trabajo son clasificación y clustering (o agrupamiento). Los algoritmos de clustering, pertenecientes a los modelos descriptivos, obtienen agrupaciones de individuos en función de sus características comunes; realizan una clasificación no supervisada. Un algoritmo de clasificación, que se puede utilizar tanto como modelo descriptivo o como modelo predictivo, estima los valores de la clase que se quiere predecir (variable clase) a partir de valores de otras variables (variables predictoras);estos algoritmos realizan una clasificación supervisada.

IV. En la fase de modelado de datos, se debe establecer una buena configuración de parámetros para la ejecución de los algoritmos empleados. Además, habrá que medir la bondad de los modelos obtenidos, para lo cual existen diferentes medidas. Por todo ello, se ha llegado a la conclusión de que los estudios estadísticos y probabilísticos son importantes para entender tanto la ejecución de los algoritmos como la interpretación correcta de los resultados. Por lo tanto, la persona que desempeñe el trabajo de analista de minería de datos, tiene que poseer dichos conocimiento, además de los requeridos como informático.

V. Cabe destacar que este proyecto no ha sido desempeñado por un profesional en el dominio del "*análisis de virus informáticos*" y tal como se ha mencionado desde el principio, tampoco se disponía de persona responsable de la fuente de datos con la que trabajar conjuntamente. Sin embargo, este hecho no ha supuesto un impedimento para el desarrollo del trabajo. Todas las decisiones que se han tomado y las acciones que se han llevado a cabo, se basan en el estudio y aplicación de las técnicas de minería de datos. Habiendo concluido el proyecto satisfactoriamente y habiendo cumplido todos los objetivos, se puede decir que el desarrollador de este trabajo se ha familiarizado con el proceso KDD , con la utilización de la metodología CRISP-DM y con la herramienta de minería de datos WEKA.

VIII. BIBLIOGRAFÍA

[1] Data Mining
http://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos

[2] Pete Chapman et al., *CRISP-DM 1.0 Step-by-step data mining guide*, 2000.
http://www.crisp-dm.org/Process/index.htm

[3] Proceso KDD
http://www.data-mining-blog.com/data-mining/data-mining-kdd-environment-fayyad-semma-five-sas-spss-crisp-dm/

[4] Usama Fayyadet al., *From Data Mining to Knowledge Discovery in Databases*, AI Magazine, Vol. 17, No. 3. 1996.

[5] Fases KDD
http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/

[6] Wikipedia CRISP-DM
http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

[7] Herramienta Weka
http://www.cs.waikato.ac.nz/ml/weka/

[8] FAQ Weka
http://weka.wikispaces.com/

[9]Jau-Hwang Wang et al.,*Virus detection using data mining techniques,*
http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1297538&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D1297538

[10] Concepción Bielza y Pedro Larrañaga, *Introducción Minería de datos*.pdf

[11] Miguel Garre, Juan José Cuadrado y Miguel Ángel Sicilia, *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software*

[12] Fases Crisp-DM
http://www.dataprix.com/

[13]Crisp-dm
http://crips-dm.wikispaces.com/1.1.+FASES

[14]Plan de contingencia
http://www.pymesycalidad20.com/como-hacer-un-plan-de-gestion-de-riesgos-para-implementar-iso-9001.html

[15] Guía de usuario de CRISP-DM
http://www.dataprix.com/es/la-gu-usuario-crisp-dm

[16] Paquete XAMPP
http://www.apachefriends.org/es/xampp.html

[17] Understanding log likelihood
http://weka.8497.n7.nabble.com/understanding-log-likelihood-in-EM-td191.html

[18] Wikipedia – Precision and Recall
http://en.wikipedia.org/wiki/Precision_and_recall

[19] Wikipedia – Algoritmo K-means
http://es.wikipedia.org/wiki/K-means

[20] K-Meansclustering en WEKA
http://facweb.cs.depaul.edu/mobasher/classes/ect584/weka/k-means.html

[21] Arboles de clasificación en WEKA
http://aprendest-022011.wikispaces.com/file/view/arboles+en+Weka.pdf

[22] Wikipedia – algoritmo EM
http://es.wikipedia.org/wiki/Algoritmo_esperanza-maximizaci%C3%B3n

[23] Ing. Corso, Cynthia Lorena,*Aplicación de algoritmos de clasificación supervisada usando Weka*,
http://www.investigacion.frc.utn.edu.ar/labsis/Publicaciones/congresos_labsis/cynthia/CNIT_2009_Aplicacion_Algoritmos_Weka.pdf

[24] Igor Baskin and Alexandre Varnek,*Tutorial onClassification*

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=Facultad de Informatica - UPM, C=ES
	Fecha/Hora	Fri Feb 14 20:13:06 CET 2014
	Emisor del Certificado	EMAILADDRESS=camanager@fi.upm.es, CN=CA Facultad de Informatica, O=Facultad de Informatica - UPM, C=ES
	Numero de Serie	630
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)